Linguistic complexity as a way to probe into genre differences?

Zarah Weiss

based on work done in collaboration with
Nina Hicks, Thomas Studer, and Detmar Meurers

Department of Linguistics & LEAD Research Network
University of Tübingen

Workshop on Complexity and Register (CAR21), Berlin, November 19th, 2021

How do we define complexity?

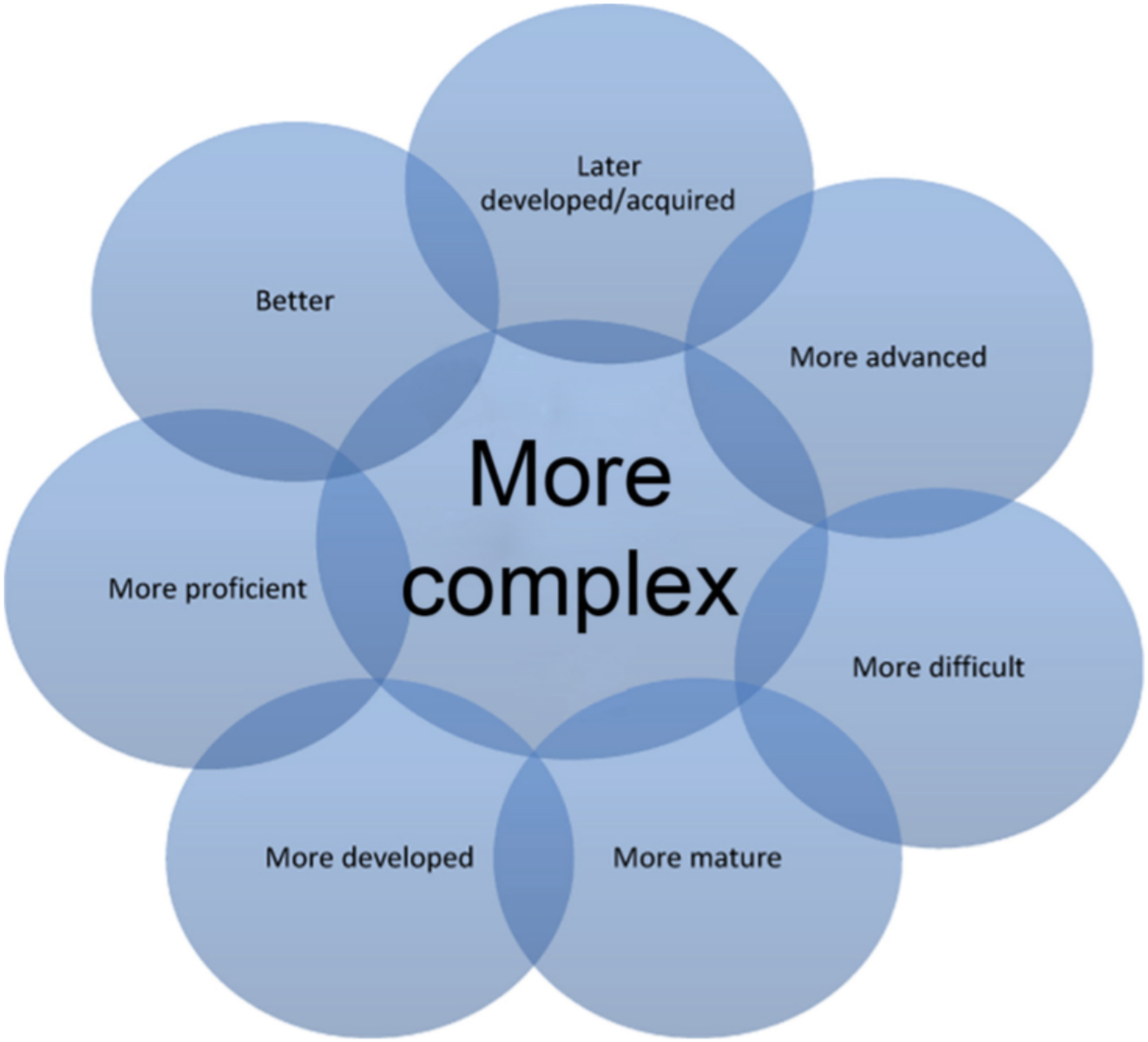


Figure from Bulté & Housen (2014)

CAF in Second Language Acquisition research

- CAF triad describes dimensions of language performance (Housen et al. 2012)
  - Assessed to characterize L2 proficiency and development
  - Also used for readability and writing quality assessment (Vajjala & Meurers 2012; Crossley et al. 2019)
- Complexity = elaborateness and variation of language
  - Accuracy = native-like and error-free language use
  - Fluency = native-like production speed and smoothness
- Adequacy suggested as 4th CAF dimension (Pallotti 2009)
  - Or: Native-like elaborateness and variation of language?

Task effects on complexity, accuracy, fluency

- Task effects have been shown to influence CAF (Yoon & Polio 2017a; Yang et al. 2015; Yoon 2017)
  - L2 corpora rarely control for comparable task conditions (Tracy-Ventura & Myles 2015; Yoon & Polio 2017a)
- Debated which task factors contribute to task effects
  - Functional task factors (e.g., genre, topic)
  - Cognitive task factors (e.g., shared context, code complexity)
- Effect of cognitive task factors on CAF trade-offs debated
  - Limited Attentional Capacity Model (Skehan 1996)
  - Cognition Hypothesis (Robinson 2001)

Task effects on complexity, accuracy, fluency

- Task effects have been shown to influence CAF (Yoon & Polio 2017a; Yang et al. 2015; Yoon 2017)
  - L2 corpora rarely control for comparable task conditions (Tracy-Ventura & Myles 2015; Yoon & Polio 2017a)
- Debated which task factors contribute to task effects
  - Functional task factors (e.g., genre, topic)
  - Cognitive task factors (e.g., shared context, code complexity)
- Effect of cognitive task factors on CAF trade-offs debated
  - Limited Attentional Capacity Model (Skehan 1996)
  - Cognition Hypothesis (Robinson 2001)

⇒ No consensus on interaction between task factors, proficiency and CAF (sub)dimensions across languages

Focus of today's talk

1. Move from focus on proficiency to focus on genre
   - ⇒ Can we use linguistic complexity to characterize genre differences in L2 writing?
   - ⇒ Does this generalize to writings by (near-)native speakers?
2. Consider multilingual dimension of genre differences
   - ⇒ Does our generalization apply across languages?
3. Zoom into informative linguistic differences between genres
   - ⇒ Which linguistic features characterize genre differences?

Automatic assessment of complexity

- Extract 489 features of German linguistic complexity, language use, human processing, and discourse cohesion
- It has been successfully used for the assessment of, e.g.,
  - Text readability and L2 proficiency (Weiss et al. 2021; Weiss & Meurers 2018, 2019b, 2021)

[1] http://ctapweb.com

Automatic assessment of complexity

- Extract 489 features of German linguistic complexity, language use, human processing, and discourse cohesion
- It has been successfully used for the assessment of, e.g.,
  - Text readability and L2 proficiency (Weiss et al. 2021; Weiss & Meurers 2018, 2019b, 2021)
  - Academic language development (Weiss & Meurers 2019a)

[1] http://ctapweb.com

Automatic assessment of complexity

- Extract 489 features of German linguistic complexity, language use, human processing, and discourse cohesion
- It has been successfully used for the assessment of, e.g.,
  - Text readability and L2 proficiency (Weiss et al. 2021; Weiss & Meurers 2018, 2019b, 2021)
  - Academic language development (Weiss & Meurers 2019a)
  - Teachers' grading behavior (Weiss et al. 2019)

[1] http://ctapweb.com

Automatic assessment of complexity

- Extract 489 features of German linguistic complexity, language use, human processing, and discourse cohesion
- It has been successfully used for the assessment of, e.g.,
  - Text readability and L2 proficiency (Weiss et al. 2021; Weiss & Meurers 2018, 2019b, 2021)
  - Academic language development (Weiss & Meurers 2019a)
  - Teachers' grading behavior (Weiss et al. 2019)
  - Spoken language in the classroom (Weiss et al. 2022)

[1] http://ctapweb.com

Automatic assessment of complexity

- Extract 489 features of German linguistic complexity, language use, human processing, and discourse cohesion
- It has been successfully used for the assessment of, e.g.,
  - Text readability and L2 proficiency (Weiss et al. 2021; Weiss & Meurers 2018, 2019b, 2021)
  - Academic language development (Weiss & Meurers 2019a)
  - Teachers' grading behavior (Weiss et al. 2019)
  - Spoken language in the classroom (Weiss et al. 2022)
- Features in multilingual Common Text Analysis Platform[1] (Chen & Meurers 2016; Weiss et al. 2021)
  - Intersection of 200+ features for 🇬🇧 🇩🇪 🇫🇷 🇪🇸
  - Soon: Italian, Arabic, Portuguese, Turkish

[1] http://ctapweb.com

Complex NPs across languages

- Hever Castle, eines der schönsten Schlösser der Welt und Wohnsitz von Anne Boleyn, hat einen geheimnisvollen Reiz.
- Hever Castle, one of the most beautiful castles in the world and the residence of Anne Boleyn, has a mysterious charm.
- Le château de Hever, l'un des plus beaux châteaux du monde et la résidence d'Anne Boleyn, a une allure mystérieuse.

Complex NPs across languages

- 🇩🇪 Hever Castle, eines der schönsten Schlösser der Welt und Wohnsitz von Anne Boleyn, hat einen geheimnisvollen Reiz.
- 🇬🇧 Hever Castle, one of the most beautiful castles in the world and the residence of Anne Boleyn, has a mysterious charm.
- 🇫🇷 Le château de Hever, l'un des plus beaux châteaux du monde et la résidence d'Anne Boleyn, a une allure mystérieuse.

Complex NPs across languages

- Hever Castle, eines der schönsten Schlösser der Welt und Wohnsitz von Anne Boleyn, hat einen geheimnisvollen Reiz.
- Hever Castle, one of the most beautiful castles in the world and the residence of Anne Boleyn, has a mysterious charm.
- Le château de Hever, l'un des plus beaux châteaux du monde et la résidence d'Anne Boleyn, a une allure mystérieuse.

Complex NPs across languages

- Hever Castle, eines der schönsten Schlösser der Welt und Wohnsitz von Anne Boleyn, hat einen geheimnisvollen Reiz.
- Hever Castle, one of the most beautiful castles in the world and the residence of Anne Boleyn, has a mysterious charm.
- Le château de Hever, l'un des plus beaux châteaux du monde et la résidence d'Anne Boleyn, a une allure mystérieuse.

Complex NPs across languages

- Hever Castle, eines der schönsten Schlösser der Welt und Wohnsitz von Anne Boleyn, hat einen geheimnisvollen Reiz.
- Hever Castle, one of the most beautiful castles in the world and the residence of Anne Boleyn, has a mysterious charm.
- Le château de Hever, l'un des plus beaux châteaux du monde et la résidence d'Anne Boleyn, a une allure mystérieuse.

Broad multilingual complexity modeling

Syntax (*N=60*)

phrasal elaboration, clausal elaboration, syntactic variation

(e.g., Chen, 2017; Kyle, 2016; Wolfe-Quintero et al., 1998)

Human processing (*N=21*)

cognitive costs, dependency lengths

(e.g., Gibson, 2002; Shain et al., 2016; Weiss 2017)

Language use (*N=28*)

frequencies, age of active use

(e.g., Chen & Meurers, 2018; Wolfe-Quintero et al., 1998)

Lexicon (*N=86*)

richness, variation, density

(e.g., Lu, 2012; McCarthy & Jarvis, 2010; Wolfe-Quintero et al., 1998)

Cohesion (*N=26*)

connectives, pronouns, grammatical transitions, **sentence overlap**

(e.g., Barzilay & Lapata, 2008; Crossley et al., 2016; Graesser et al., 2004; Todirascu, 2013)

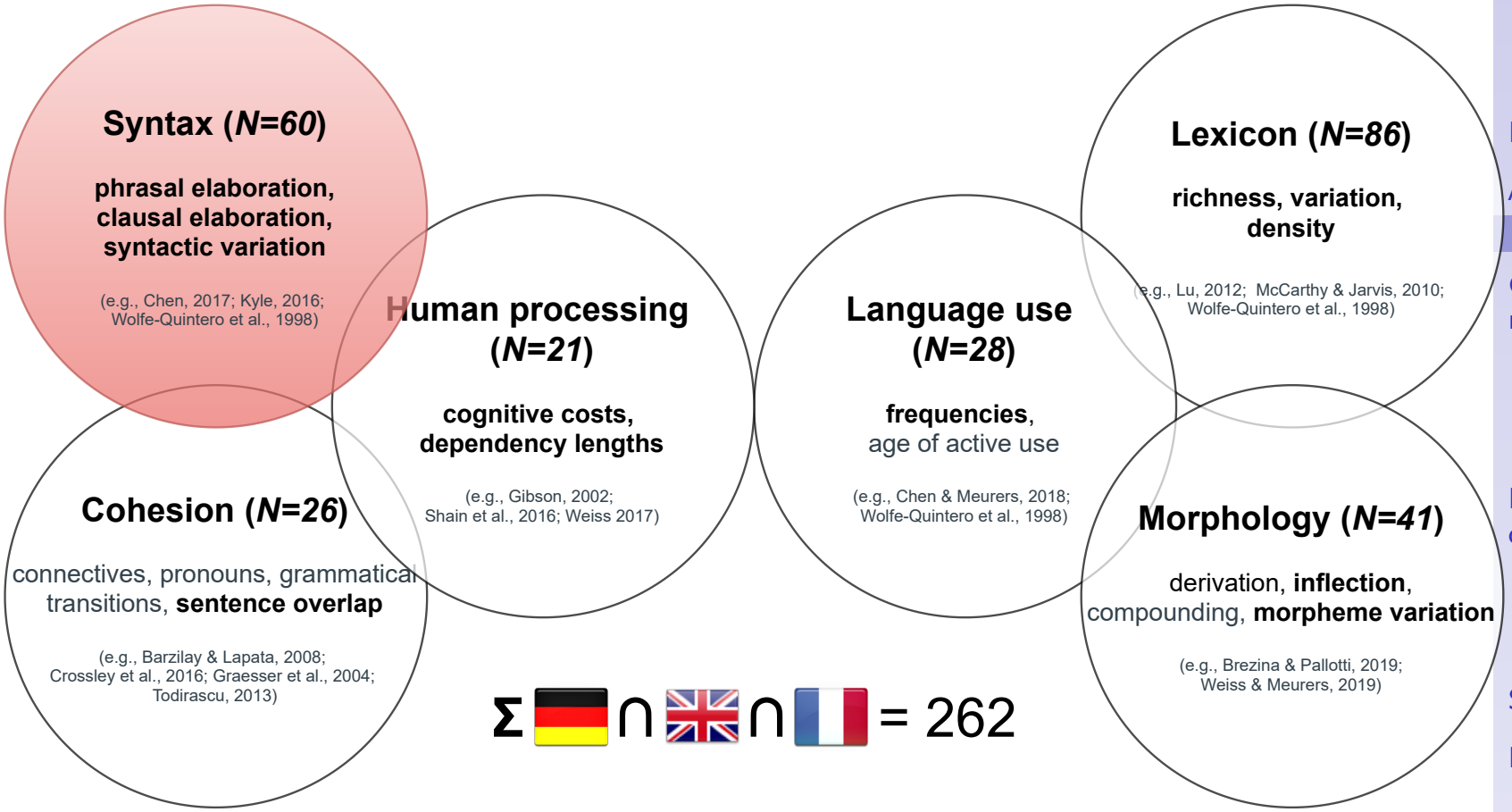
Morphology (*N=41*)

derivation, **inflection**, compounding, **morpheme variation**

(e.g., Brezina & Pallotti, 2019; Weiss & Meurers, 2019)

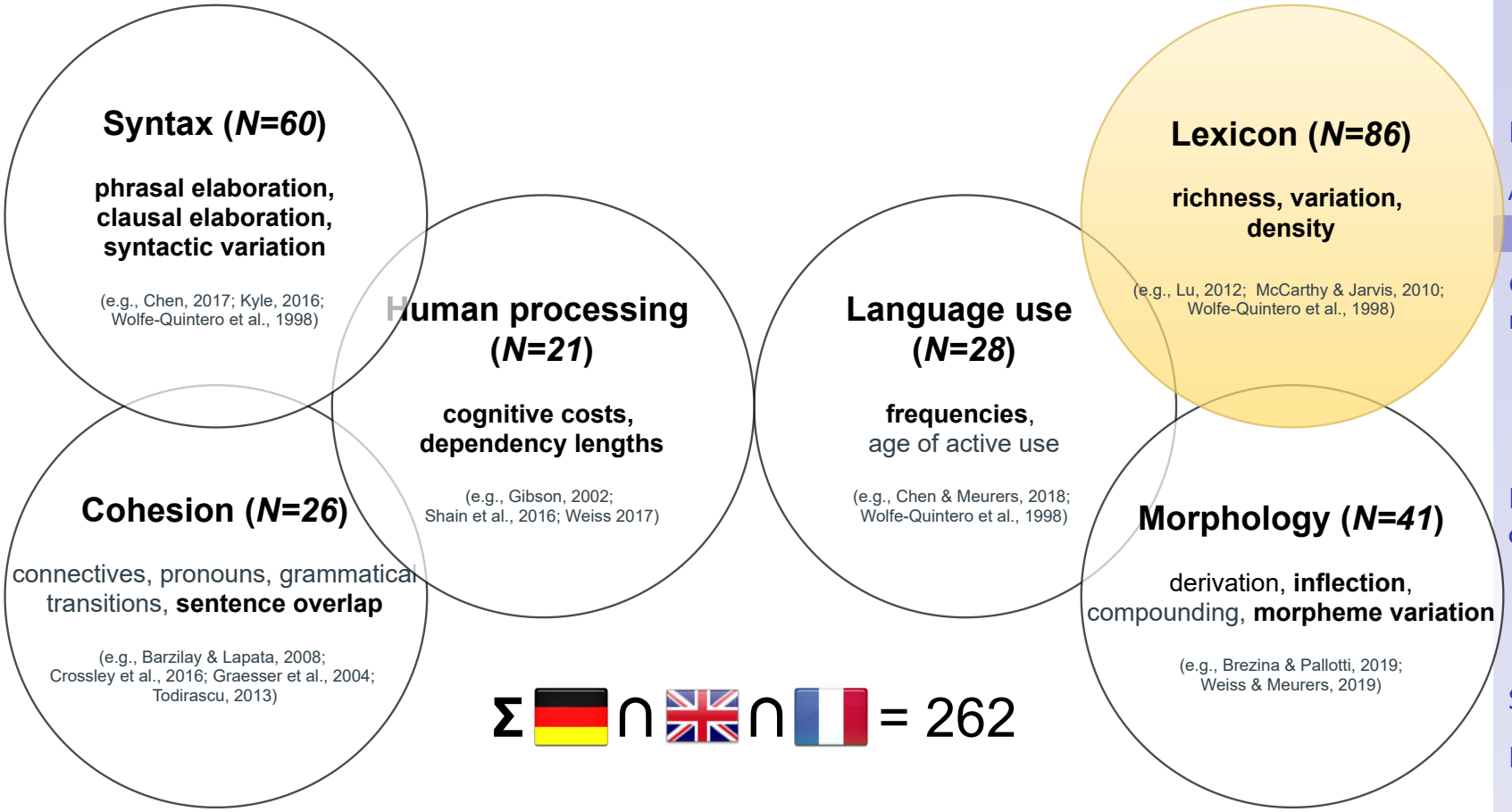
Σ 🇩🇪 $\cap$ 🇬🇧 $\cap$ 🇫🇷 = 262

Broad multilingual complexity modeling



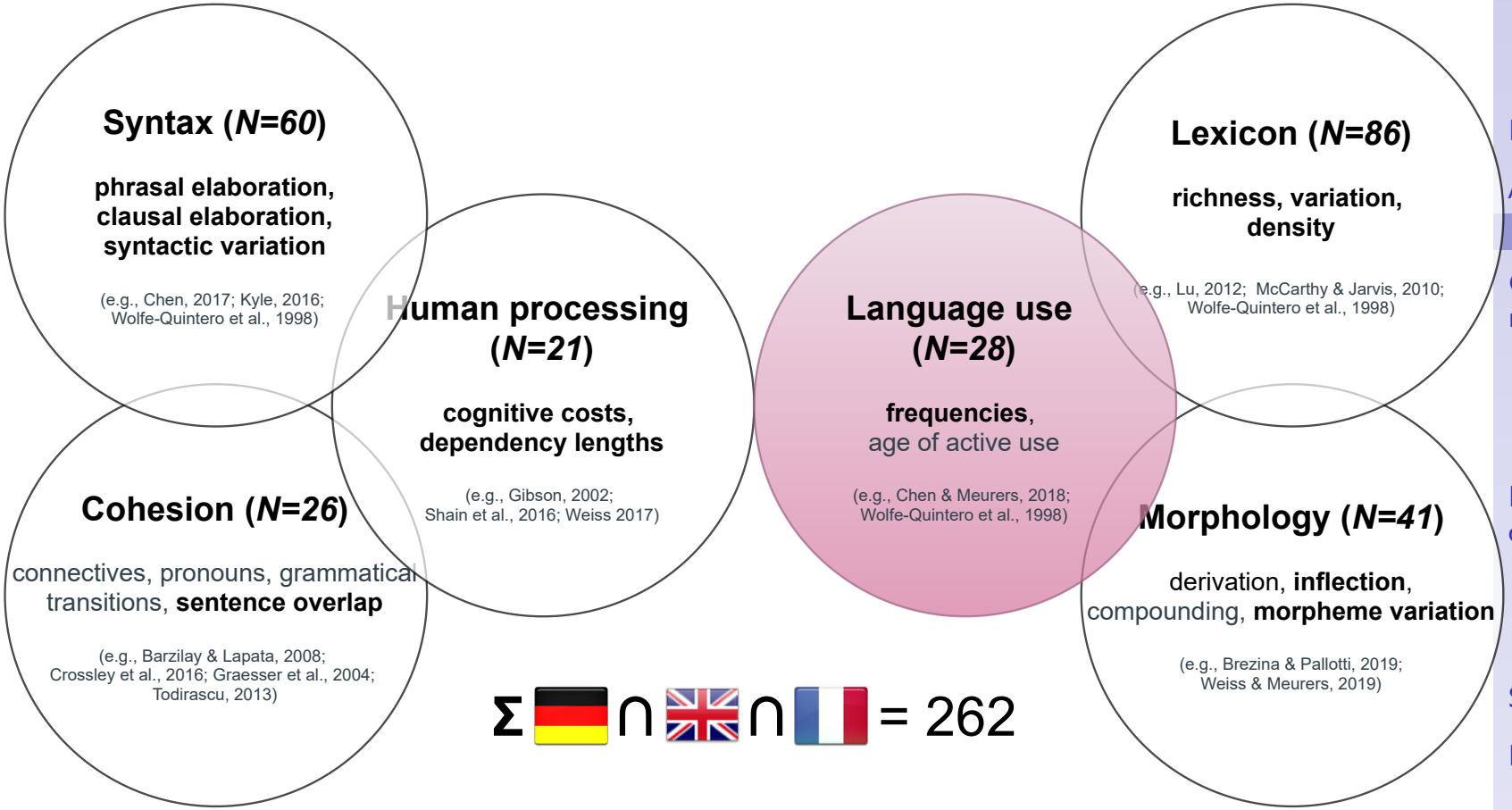
- Clausal elaboration (dependent clauses per sentence)
- Phrasal elaboration (modifiers per noun phrase)
- Syntactic variation (parse tree edit distances)

Broad multilingual complexity modeling



- Richness (type-token ratios)
- Variation (POS-specific type-token ratios)
- Density (noun density = nouns per lexical word)

Broad multilingual complexity modeling



Σ 🇩🇪 $\cap$ 🇬🇧 $\cap$ 🇫🇷 = 262

- Mean and SD word frequency from frequency data bases
- Here: frequencies per million words for TV captions
- Separate for different POS or word categories

Broad multilingual complexity modeling

Syntax (*N=60*)

phrasal elaboration, clausal elaboration, syntactic variation

(e.g., Chen, 2017; Kyle, 2016; Wolfe-Quintero et al., 1998)

Human processing (*N=21*)

cognitive costs, dependency lengths

(e.g., Gibson, 2002; Shain et al., 2016; Weiss 2017)

Language use (*N=28*)

frequencies, age of active use

(e.g., Chen & Meurers, 2018; Wolfe-Quintero et al., 1998)

Lexicon (*N=86*)

richness, variation, density

(e.g., Lu, 2012; McCarthy & Jarvis, 2010; Wolfe-Quintero et al., 1998)

Cohesion (*N=26*)

connectives, pronouns, grammatical transitions, **sentence overlap**

(e.g., Barzilay & Lapata, 2008; Crossley et al., 2016; Graesser et al., 2004; Todirascu, 2013)

Morphology (*N=41*)

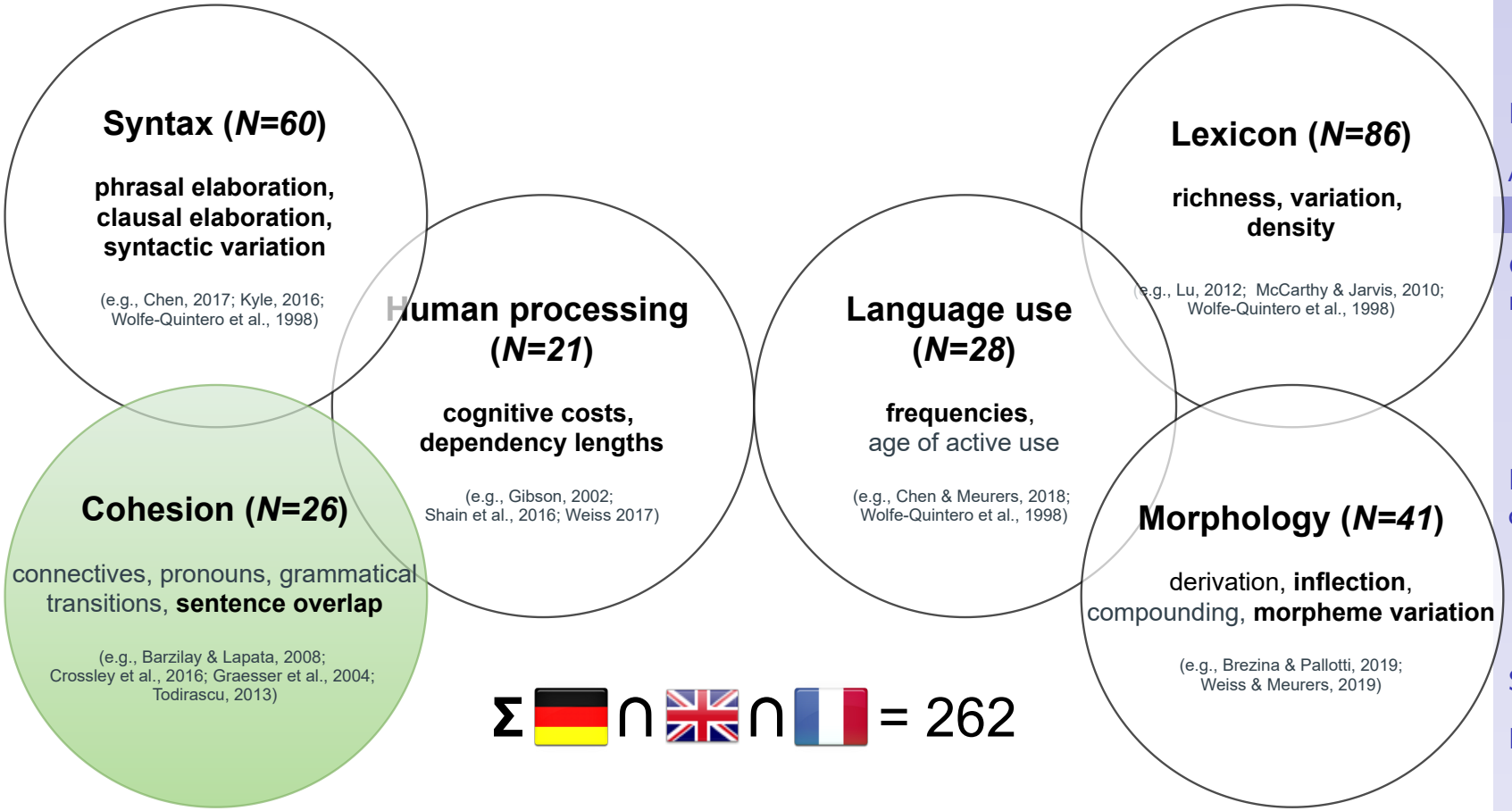
derivation, **inflection**, compounding, **morpheme variation**

(e.g., Brezina & Pallotti, 2019; Weiss & Meurers, 2019)

Σ 🇩🇪 $\cap$ 🇬🇧 $\cap$ 🇫🇷 = 262

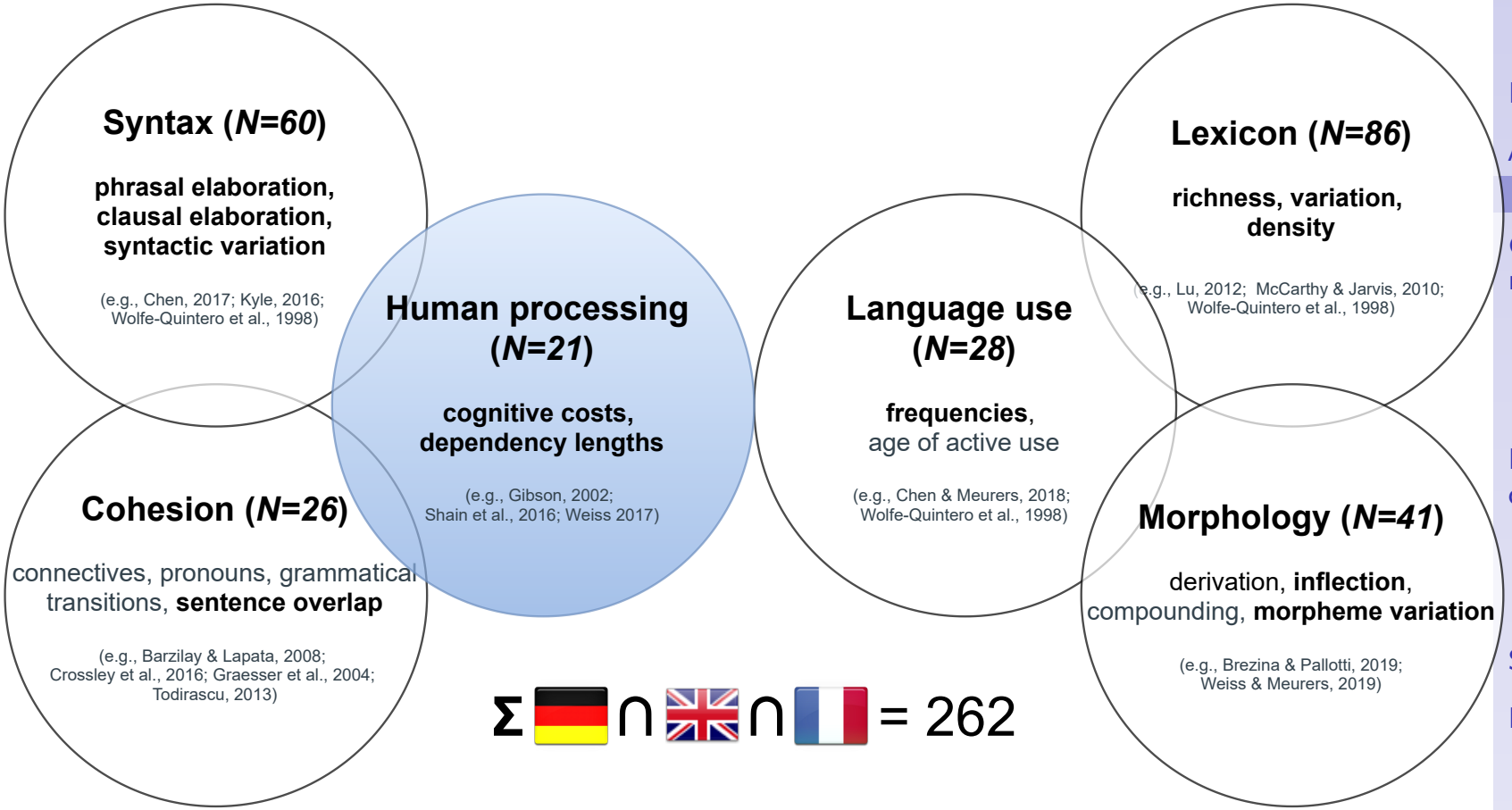
- Morphological complexity index (Brezina & Pallotti 2019) to measure variability of morphemes for specific POS

Broad multilingual complexity modeling



- Overlap of POS/arguments between
  1. Adjacent sentences = local overlap
  2. Any two sentences = global overlap

Broad multilingual complexity modeling



- Average and maximal integration costs based on Dependency Locality Theory by Gibson (2000)

Multiligual complexity assessment

1. Shared NLP pipeline to obtain sentences, words, POS tags, constituency/dependency parses, lemmas, stems (see for technical details Weiss et al. 2021)
2. Identify linguistic constructions with extraction rules
   - Rules for language-specific features
   - Frequency features from Subtlex-US/DE and Lexique (Brysbaert et al. 2011a,b; New et al. 2001)
   - Language-specific rules for syllables, POS, constituents
   - Language-independent rules for all other
3. Fully language-independent calculation of feature ratios

Cross-lingual readability assessment

- Predict reading levels on multi-lingual L2 reading corpus (Weiss et al. 2021)
  1. Can we predict L2 reading levels for German and English?
  2. Does this generalize across languages?
  3. How do reading levels linguistically differ from each other

[2] www.spotlight-online.de, www.deutsch-perfekt.de

Cross-lingual readability assessment

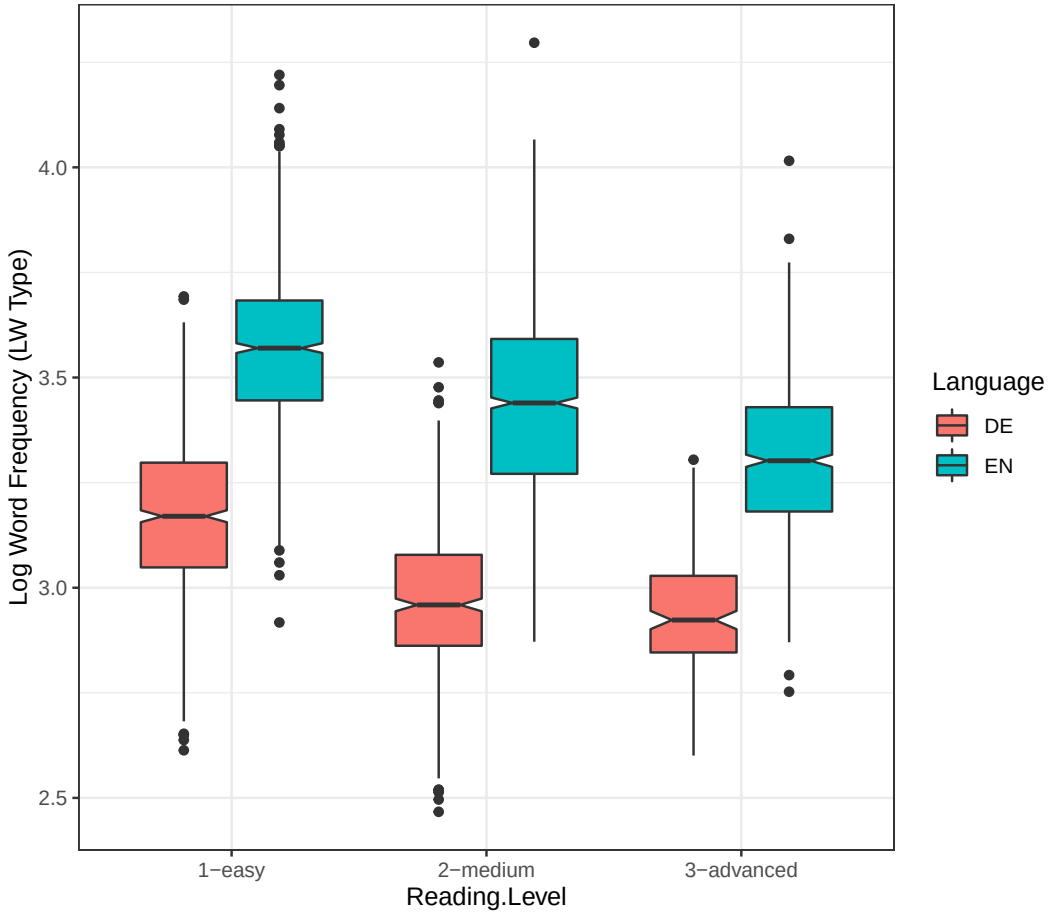
- Predict reading levels on multi-lingual L2 reading corpus (Weiss et al. 2021)
  1. Can we predict L2 reading levels for German and English?
  2. Does this generalize across languages?
  3. How do reading levels linguistically differ from each other
- Experiment on Spotlight corpus from L2 magazine articles[2]
  - English (N=3.285), German (N=1.446)
  - Beginning (A2), medium (B1/B2), advanced (C1)

[2] www.spotlight-online.de, www.deutsch-perfekt.de

Readability classification results on Spotlight

- Train/test ordinal random forest using 301 features
  - Model 1: distinguish reading levels on English data
  - Model 2: distinguish reading levels on German data
- Evaluation with accuracy in with 10-fold cross-validation
  - Model 1: 74.5% for English; baseline: 46.5%
  - Model 2: 88.0% for German; baseline: 52.8%
- Language-specific models generalize to some degree to the other language (despite considerable performance drop)
- Investigation of informative features show comparable reading level differences
  - Especially wrt. lexical domain and frequency features

Illustrate feature differences for English/German



- More frequent vocabulary in English than German texts
  - Likely due to compounding and inflection in German
- Word frequency decreases with increasing reading levels

Proficiency and task effects

- Predict German L2 proficiency of learners from their written essays (Weiss & Meurers 2019b)
  1. Can we predict proficiency from learner writing?
  2. Which task influences do we observe?
  3. How do proficiency, task, and complexity interact?

Proficiency and task effects

- Predict German L2 proficiency of learners from their written essays (Weiss & Meurers 2019b)
  1. Can we predict proficiency from learner writing?
  2. Which task influences do we observe?
  3. How do proficiency, task, and complexity interact?
- Merlin corpus of 1,033 German L2 texts (Wisniewski et al. 2013)
  - Elicited in standardized testing at CEFR levels A1 to C1
  - Three task prompts per level ⇒ 15 different task prompt
  - Texts rated by two experts on CEFR scale (A1 to C2)
- Focus on task rather than individual task factors

  ⇒ Nearly all factors fully confounded with test level

Proficiency and task classification results

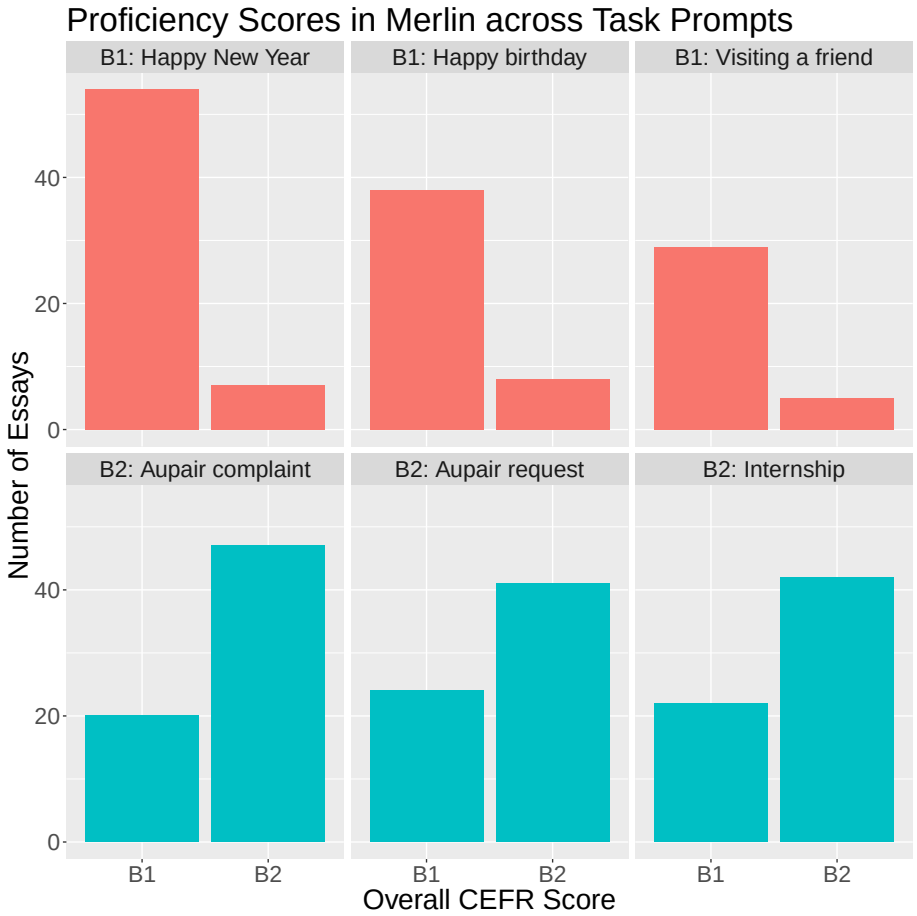
- Train/test ordinal random forests
  - 489 features extracted from form-based target hypothesis
  - Model 1: predict L2 proficiency of learners
  - Model 2: predict task prompt for text
- Evaluation with accuracy in 10-fold cross-validation
  - Model 1: 70.0%; 5-way classification; baseline: 32.0%
  - Model 2: 89.8%; 15-way classification; baseline: 8.0%

Proficiency and task classification results

- Train/test ordinal random forests
  - 489 features extracted from form-based target hypothesis
  - Model 1: predict L2 proficiency of learners
  - Model 2: predict task prompt for text
- Evaluation with accuracy in 10-fold cross-validation
  - Model 1: 70.0%; 5-way classification; baseline: 32.0%
  - Model 2: 89.8%; 15-way classification; baseline: 8.0%

⇒ Complexity features are more indicative for task differences than for proficiency differences

Zooming in on the Merlin corpus



- Subset of the Merlin corpus used for classification studies
  - Focus on B1/B2 learners taking B1/B2 tests (N=337)
- Test level confounds with genre, formality, & structured-ness

Summary of findings

- Task differences play central role in learners' L2 complexity and partially outweighs seeming proficiency differences

Summary of findings

- Task differences play central role in learners' L2 complexity and partially outweighs seeming proficiency differences
  - Morphological complexity, expl. cohesion, language use

Summary of findings

- Task differences play central role in learners' L2 complexity and partially outweighs seeming proficiency differences
  - Morphological complexity, expl. cohesion, language use
  - Lexical and phrasal complexity characterize proficiency differences despite task sensitivity

Summary of findings

- Task differences play central role in learners' L2 complexity and partially outweighs seeming proficiency differences
  - Morphological complexity, expl. cohesion, language use
  - Lexical and phrasal complexity characterize proficiency differences despite task sensitivity
  - Human processing and clausal measures stable

Summary of findings

- Task differences play central role in learners' L2 complexity and partially outweighs seeming proficiency differences
  - Morphological complexity, expl. cohesion, language use
  - Lexical and phrasal complexity characterize proficiency differences despite task sensitivity
  - Human processing and clausal measures stable
- Repeated observation of "test level" effect on complexity
  - Confounds with genre, formality, & structured-ness

Summary of findings

- Task differences play central role in learners' L2 complexity and partially outweighs seeming proficiency differences
  - Morphological complexity, expl. cohesion, language use
  - Lexical and phrasal complexity characterize proficiency differences despite task sensitivity
  - Human processing and clausal measures stable
- Repeated observation of "test level" effect on complexity
  - Confounds with genre, formality, & structured-ness
- Similar findings on Falko GT corpus (Siemen et al. 2006)
  - Corpus of German L2 writing elicited in curricular writing courses at the Georgetown University
  - No deconstruction in task characteristics possible

Complexity and genre

- Previous research demonstrates interaction between tasks, proficiency and complexity domains
- Confounded task variables in corpora do not allow to distinguish the source of these task effect
  - Are the observations really due to overall task difference or can we identify specific contribution of certain task factors?
  - Do findings generalize to L1 writing and across languages?

Complexity and genre

- Previous research demonstrates interaction between tasks, proficiency and complexity domains
- Confounded task variables in corpora do not allow to distinguish the source of these task effect
  - Are the observations really due to overall task difference or can we identify specific contribution of certain task factors?
  - Do findings generalize to L1 writing and across languages?

$\Rightarrow$ Probe into genre differences in L2 writing across languages

SWIKO corpus

- Multilingual task-based Swiss learner corpus of 410 oral & 1,803 written texts by lower secondary school students[3]
  - Currently augmented and further analyzed in two projects (WETLAND, DiCoi; 2021–2024)

[3] https://centre-plurilinguisme.ch/en/research/swiss-learner-corpus-swiko

SWIKO corpus

- Multilingual task-based Swiss learner corpus of 410 oral & 1,803 written texts by lower secondary school students[3]
  - Currently augmented and further analyzed in two projects (WETLAND, DiCoi; 2021–2024)
- Analysis of two sub-corpora of Swiss student texts
  - 1,002 L2 texts: 176🇩🇪+329🇫🇷+497🇬🇧
  - 585 in language of schooling (LoS): 332🇩🇪+152🇫🇷+101🇬🇧

[3] https://centre-plurilinguisme.ch/en/research/swiss-learner-corpus-swiko

SWIKO corpus

- Multilingual task-based Swiss learner corpus of 410 oral & 1,803 written texts by lower secondary school students[3]
  - Currently augmented and further analyzed in two projects (WETLAND, DiCoi; 2021–2024)
- Analysis of two sub-corpora of Swiss student texts
  - 1,002 L2 texts: 176🇩🇪+329🇫🇷+497🇬🇧
  - 585 in language of schooling (LoS): 332🇩🇪+152🇫🇷+101🇬🇧

[3] https://centre-plurilinguisme.ch/en/research/swiss-learner-corpus-swiko

SWIKO corpus

- Multilingual task-based Swiss learner corpus of 410 oral & 1,803 written texts by lower secondary school students[3]
  - Currently augmented and further analyzed in two projects (WETLAND, DiCoi; 2021–2024)
- Analysis of two sub-corpora of Swiss student texts
  - 1,002 L2 texts: 176🇩🇪+329🇫🇷+497🇬🇧
  - 585 in language of schooling (LoS): 332🇩🇪+152🇫🇷+101🇬🇧

[3] https://centre-plurilinguisme.ch/en/research/swiss-learner-corpus-swiko

SWIKO corpus

- Multilingual task-based Swiss learner corpus of 410 oral & 1,803 written texts by lower secondary school students[3]
  - Currently augmented and further analyzed in two projects (WETLAND, DiCoi; 2021–2024)
- Analysis of two sub-corpora of Swiss student texts
  - 1,002 L2 texts: 176🇩🇪+329🇫🇷+497🇬🇧
  - 585 in language of schooling (LoS): 332🇩🇪+152🇫🇷+101🇬🇧
- Corpus characteristics
  - Text length across languages & tasks: $\mu = 65(\pm 34)$ words
  - Annotated with form-based, local normalization
  - Systematically varies genre, formality, structured-ness

[3] https://centre-plurilinguisme.ch/en/research/swiss-learner-corpus-swiko

SWIKO tasks used to elicit data across languages

		01	02	03	04	05	06	07	08
Genre	arg.			✗	✗			✗	✗
	des.	✗	✗			✗	✗		
Formality	form.		✗		✗		✗		✗
	inform.	✗		✗		✗		✗	
Structured	yes	✗	✗	✗	✗				
	no					✗	✗	✗	✗

SWIKO tasks used to elicit data across languages

		01	02	03	04	05	06	07	08
Genre	arg.			✗	✗			✗	✗
	des.	✗	✗			✗	✗		
Formality	form.		✗		✗		✗		✗
	inform.	✗		✗		✗		✗	
Structured	yes	✗	✗	✗	✗				
	no					✗	✗	✗	✗

01 answer 5 out of 12 personal questions
(e.g., what is your favorite animal/book/food?)

SWIKO tasks used to elicit data across languages

		01	02	03	04	05	06	07	08
Genre	arg.			✗	✗			✗	✗
	des.	✗	✗			✗	✗		
Formality	form.		✗		✗		✗		✗
	inform.	✗		✗		✗		✗	
Structured	yes	✗	✗	✗	✗				
	no					✗	✗	✗	✗

02 describe four graphics about animals in Switzerland

SWIKO tasks used to elicit data across languages

		01	02	03	04	05	06	07	08
Genre	arg.			✗	✗			✗	✗
	des.	✗	✗			✗	✗		
Formality	form.		✗		✗		✗		✗
	inform.	✗		✗		✗		✗	
Structured	yes	✗	✗	✗	✗				
	no					✗	✗	✗	✗

03 discuss popularity of vacation destinations from a list

SWIKO tasks used to elicit data across languages

		01	02	03	04	05	06	07	08
Genre	arg.			✗	✗			✗	✗
	des.	✗	✗			✗	✗		
Formality	form.		✗		✗		✗		✗
	inform.	✗		✗		✗		✗	
Structured	yes	✗	✗	✗	✗				
	no					✗	✗	✗	✗

04 discuss list of top 1–5, 95–100 most important inventions

SWIKO tasks used to elicit data across languages

		01	02	03	04	05	06	07	08
Genre	arg.			✗	✗			✗	✗
	des.	✗	✗			✗	✗		
Formality	form.		✗		✗		✗		✗
	inform.	✗		✗		✗		✗	
Structured	yes	✗	✗	✗	✗				
	no					✗	✗	✗	✗

05 write free self-description for a school exchange

SWIKO tasks used to elicit data across languages

		01	02	03	04	05	06	07	08
Genre	arg.			✗	✗			✗	✗
	des.	✗	✗			✗	✗		
Formality	form.		✗		✗		✗		✗
	inform.	✗		✗		✗		✗	
Structured	yes	✗	✗	✗	✗				
	no					✗	✗	✗	✗

06 briefly present one of the presented topics (8 options)

SWIKO tasks used to elicit data across languages

		01	02	03	04	05	06	07	08
Genre	arg.			✗	✗			✗	✗
	des.	✗	✗			✗	✗		
Formality	form.		✗		✗		✗		✗
	inform.	✗		✗		✗		✗	
Structured	yes	✗	✗	✗	✗				
	no					✗	✗	✗	✗

07 discuss if school should start and end later (10am–5:30pm)

SWIKO tasks used to elicit data across languages

		01	02	03	04	05	06	07	08
Genre	arg.			✗	✗			✗	✗
	des.	✗	✗			✗	✗		
Formality	form.		✗		✗		✗		✗
	inform.	✗		✗		✗		✗	
Structured	yes	✗	✗	✗	✗				
	no					✗	✗	✗	✗

08 discuss exchanging foreign language classes in school against obligatory language-exchange for six months

Classification experiment

- Predict text genre (arg./des.) from L2 writing in 🇩🇪 🇫🇷 🇬🇧
  - Can we train a model that generalizes across all languages?
  - Does the model generalize form L2 to LoS?

Classification experiment

- Predict text genre (arg./des.) from L2 writing in 🇩🇪 🇫🇷 🇬🇧
  - Can we train a model that generalizes across all languages?
  - Does the model generalize form L2 to LoS?
- Set-up for random forest classifier (best of 5 algorithms)
  1. Use variable features available for 🇩🇪 🇫🇷 🇬🇧: 200/262
  2. Train/test on L2 data with 10-fold cross-validation
  3. Apply model trained on L2 on LoS data
- Classifier performance in accuracy
  - Across folds: 80.4%[79.6; 81.3] (baseline: 51.1%)
  - On LoS data: 84.3% (baseline: 50.1%)

  $\Rightarrow$ Clear improvement from L2 to LoS texts

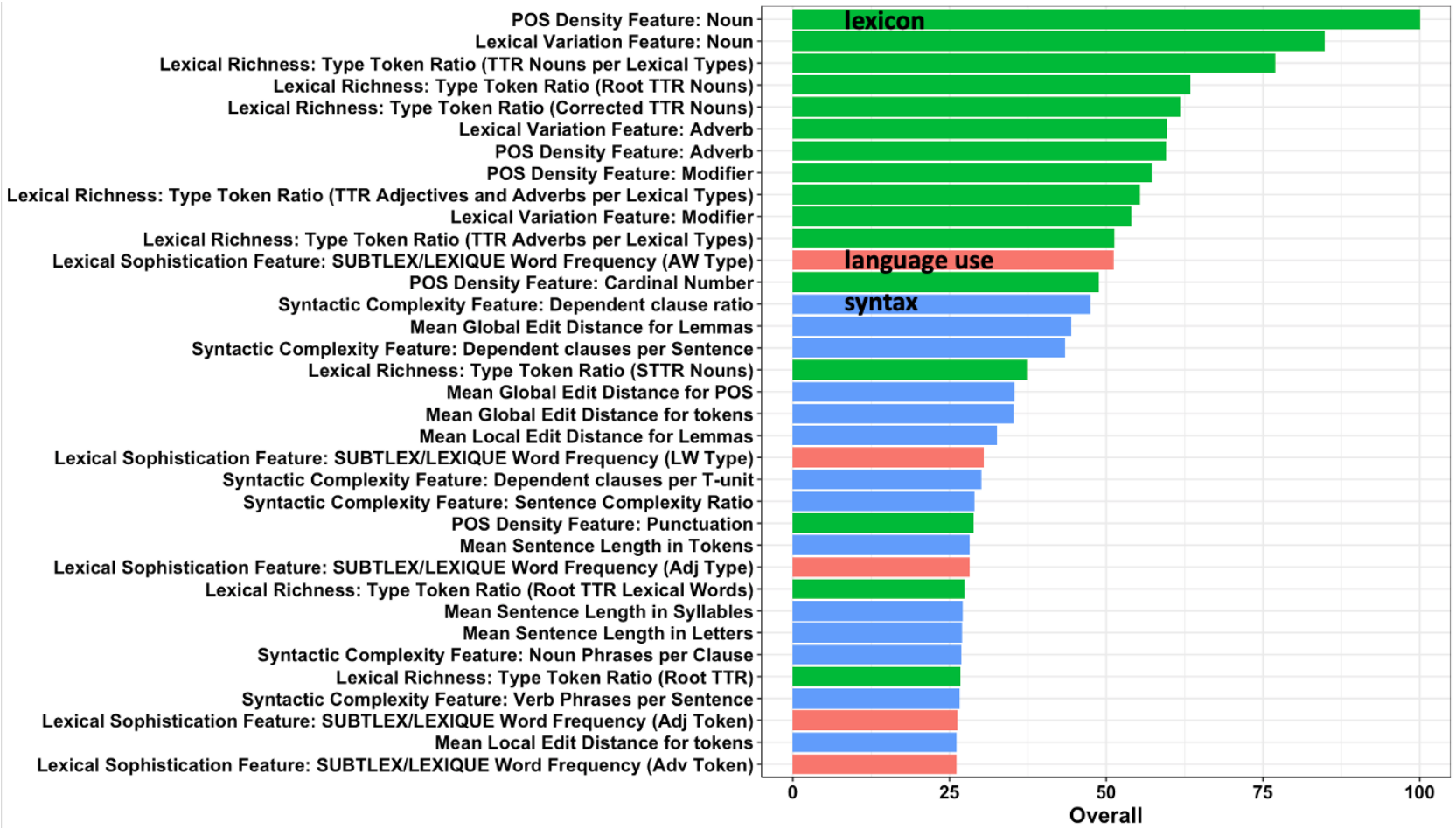
Discussion

1. Can we train a model that generalizes across all languages?
   - Yes, our model successfully learns linguistic genre characteristics that generalize across languages
2. Does the model generalize form L2 to language of schooling?
   - Yes, a genre model trained on L2 data generalizes to language of schooling data
   - Higher performance could indicate that linguistic genre differences are more pronounced in LoS data

Which linguistic text differences inform the model?

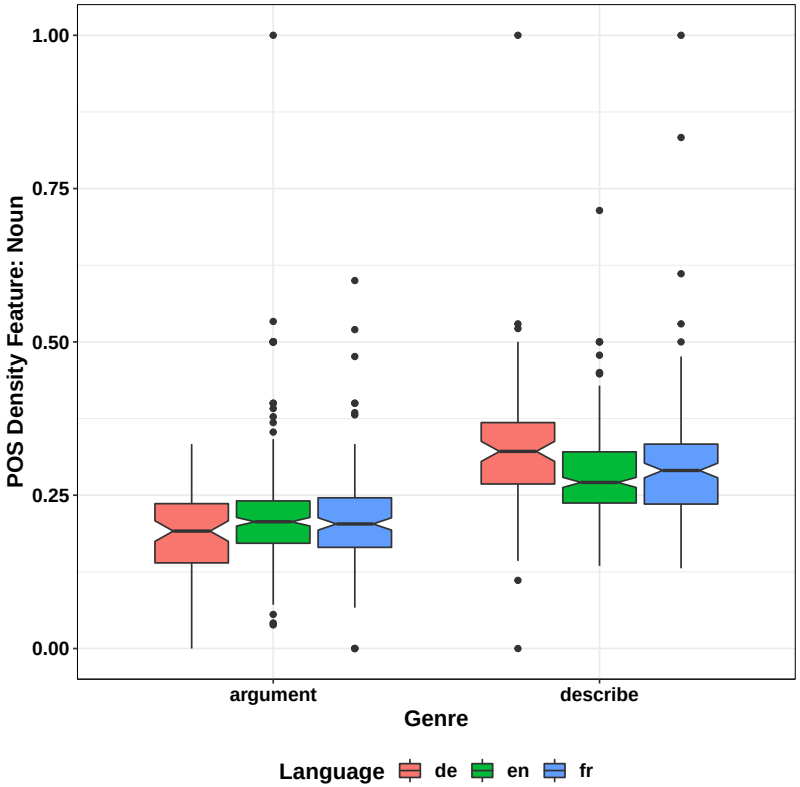
- Investigate variable importance of features in model
  - Compares avg. accuracy difference with and w/o each feature
  - Normalized with standard error
- Compare top & bottom 35 features grouped by domain & features commonly associated with argumentative texts

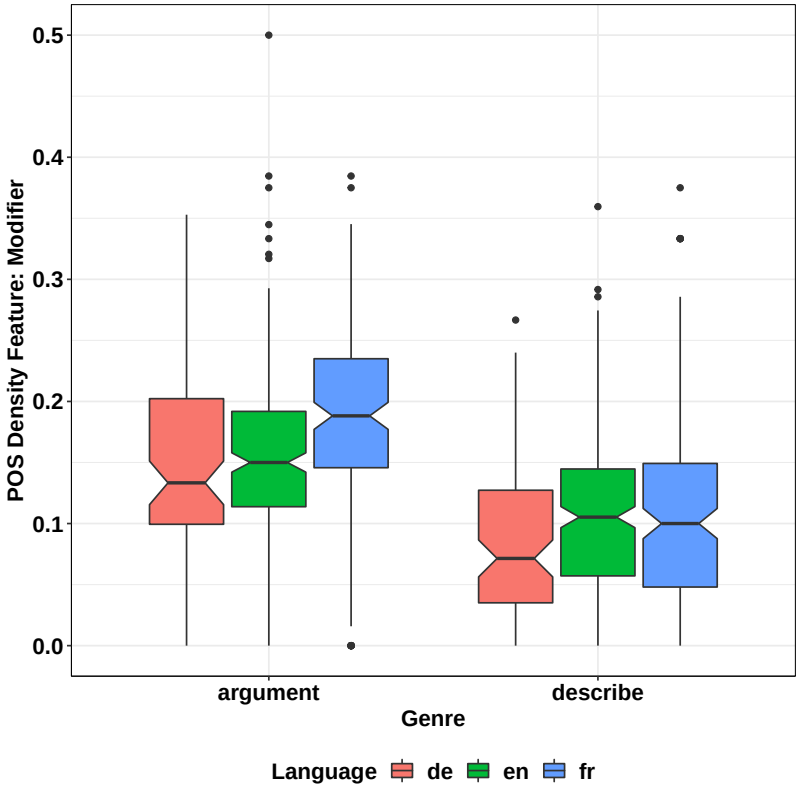
Most informative features



- Lexical measures of noun and modifier density and variation
- Clausal elaboration, syntactic variation and word frequency

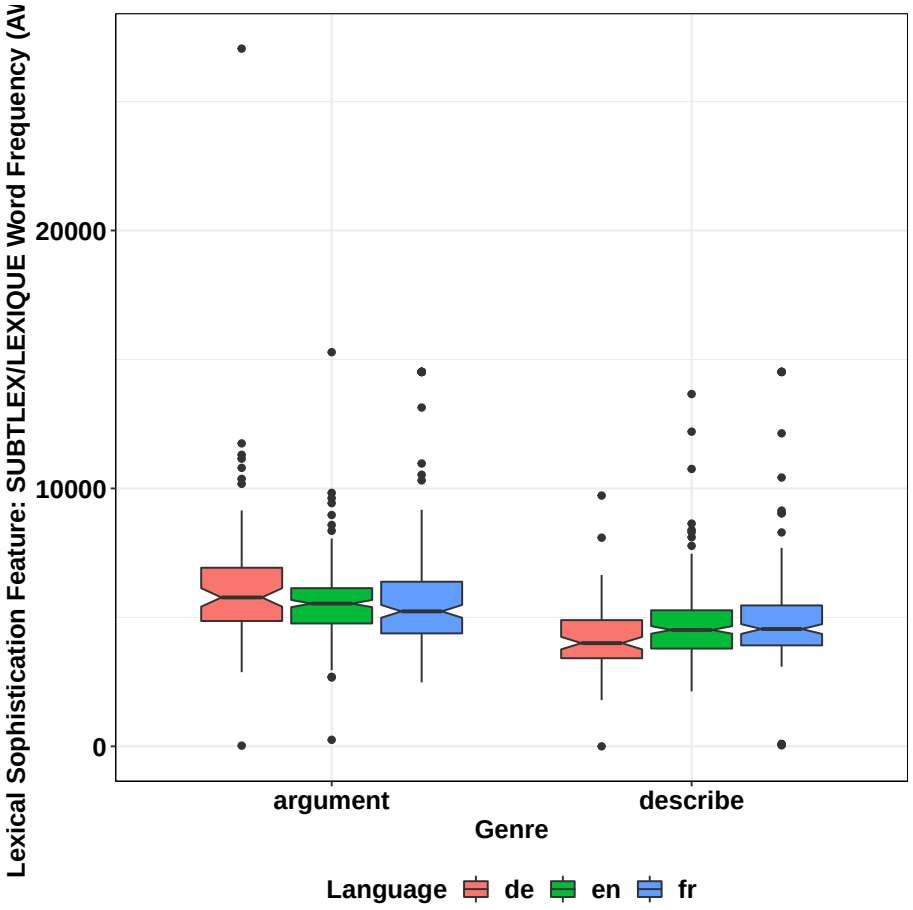
Noun and modifier density





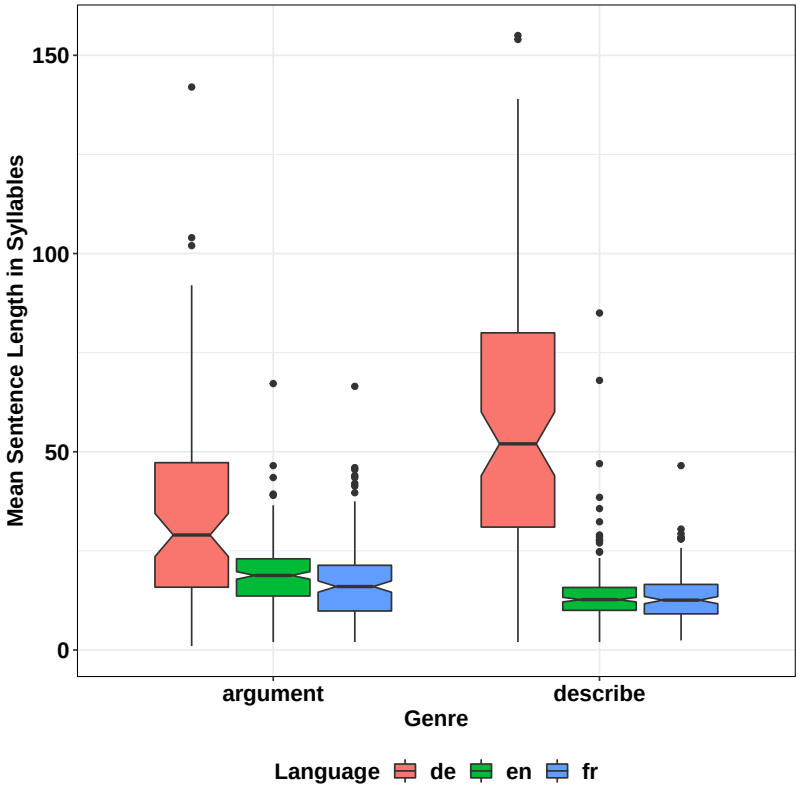
- Descriptive texts show higher noun & lower modifier density than argumentative texts
  - Also representative for noun and modifier variation features

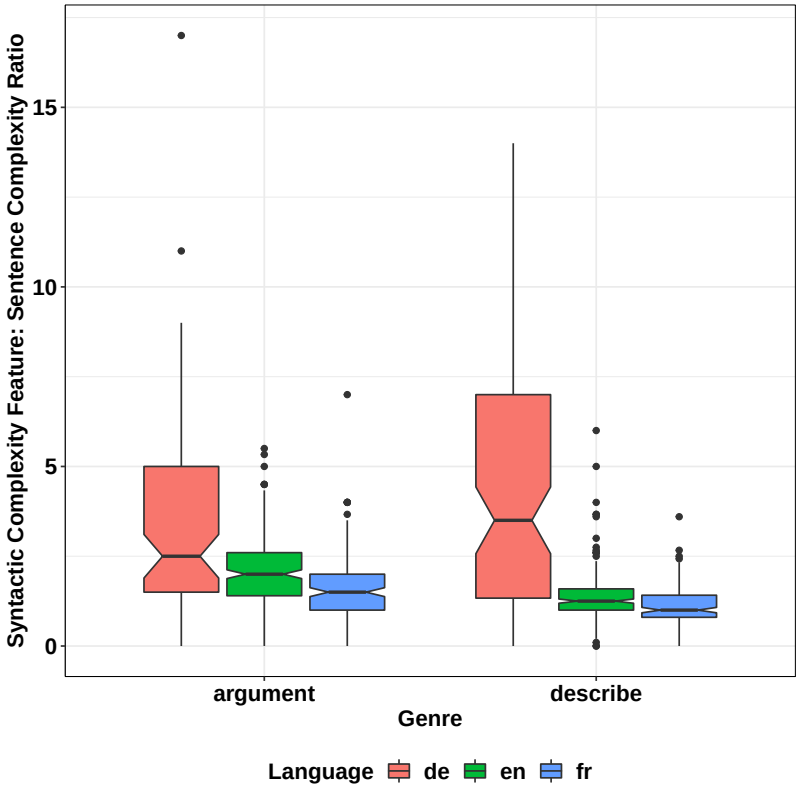
Word frequencies



- Descriptive texts contain less frequent vocabulary than argumentative texts (across all languages)

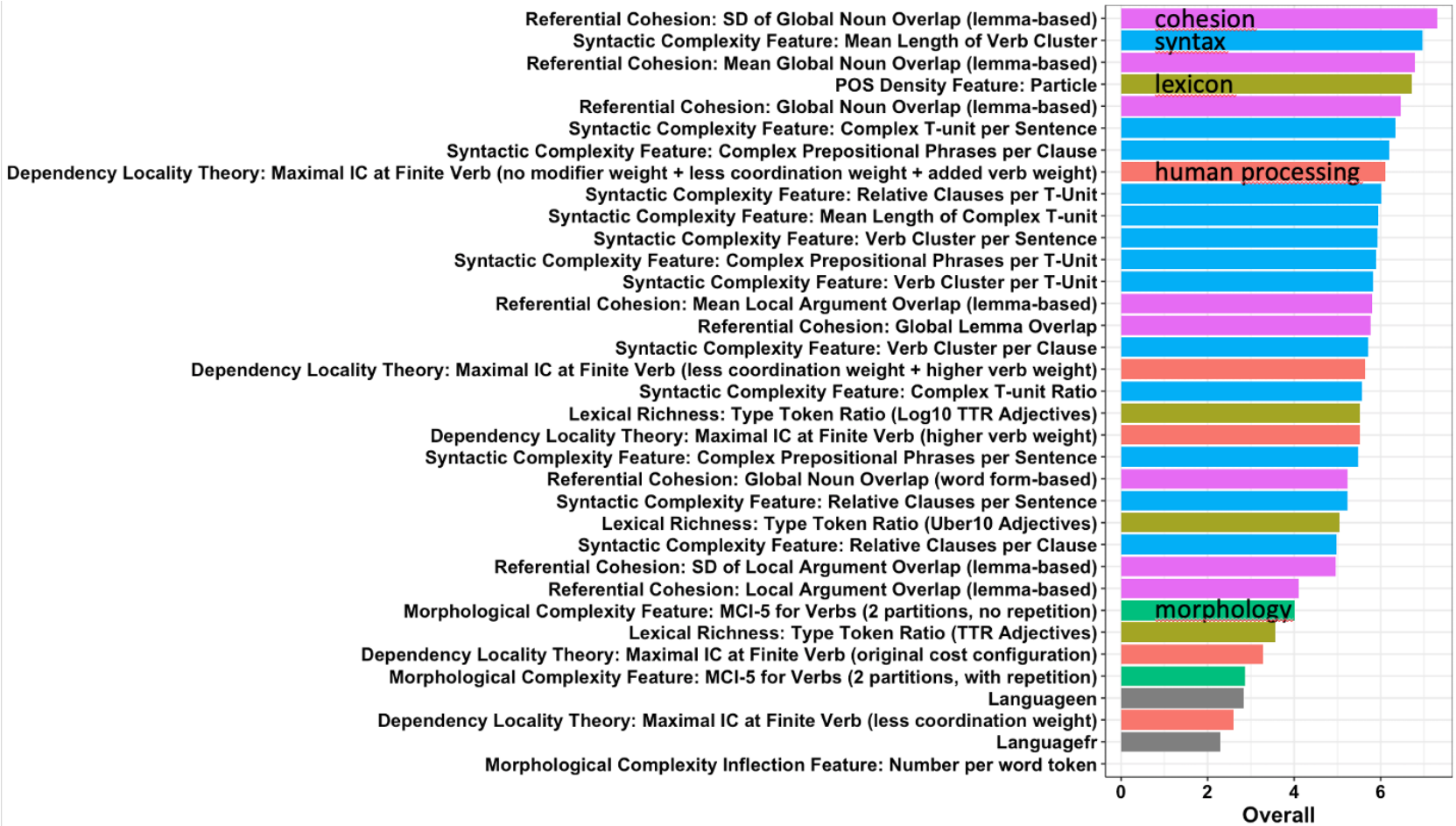
Clausal elaboration





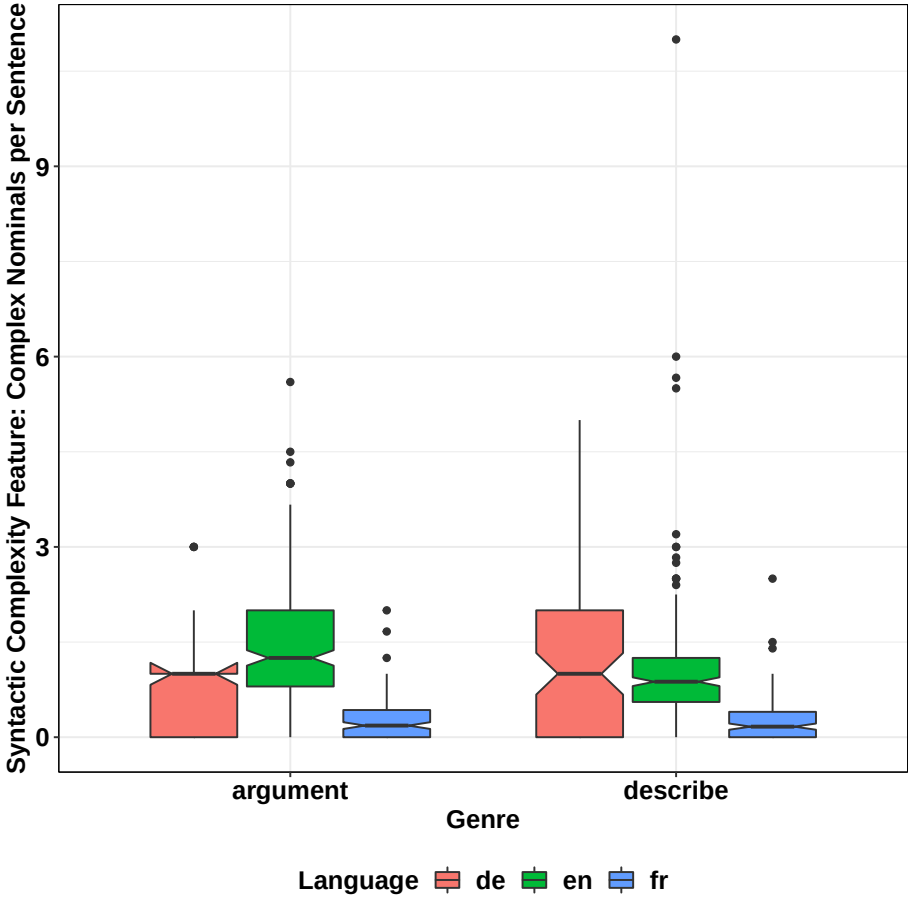
- Argumentative texts contain longer, more elaborate clauses
  - Matches contrast argumentative vs. narrative texts in English L2 writing (Yoon & Polio 2017a,b; Biber & Conrad 2019)
- Effect only for French and English, inverse for German

Least informative features



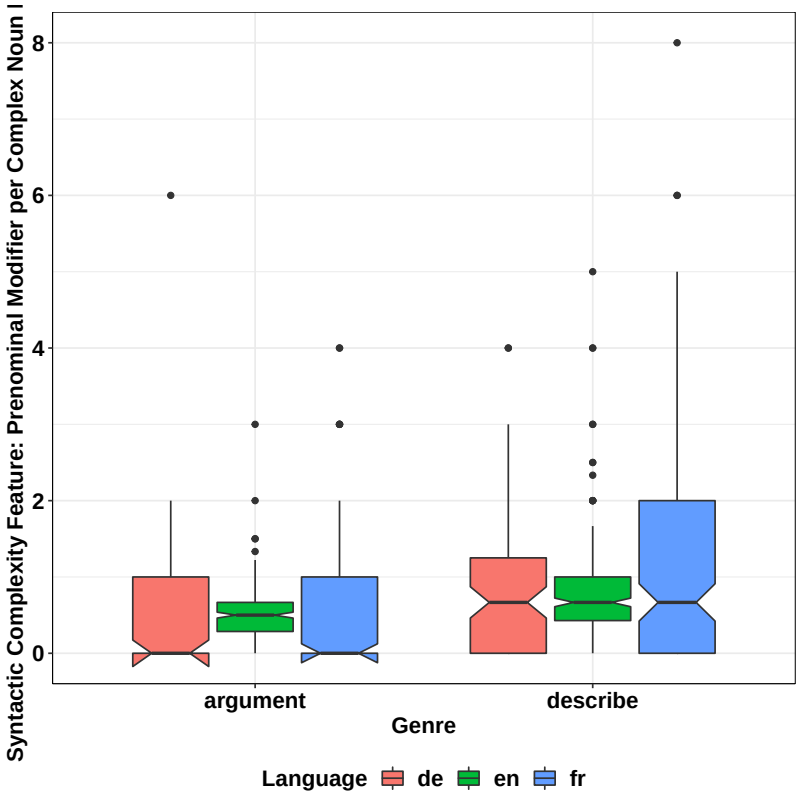
- Implicit cohesion and human processing measures
  - Partially in line with findings on Merlin/Falko
- Other: relative clauses & PP/VP complexity

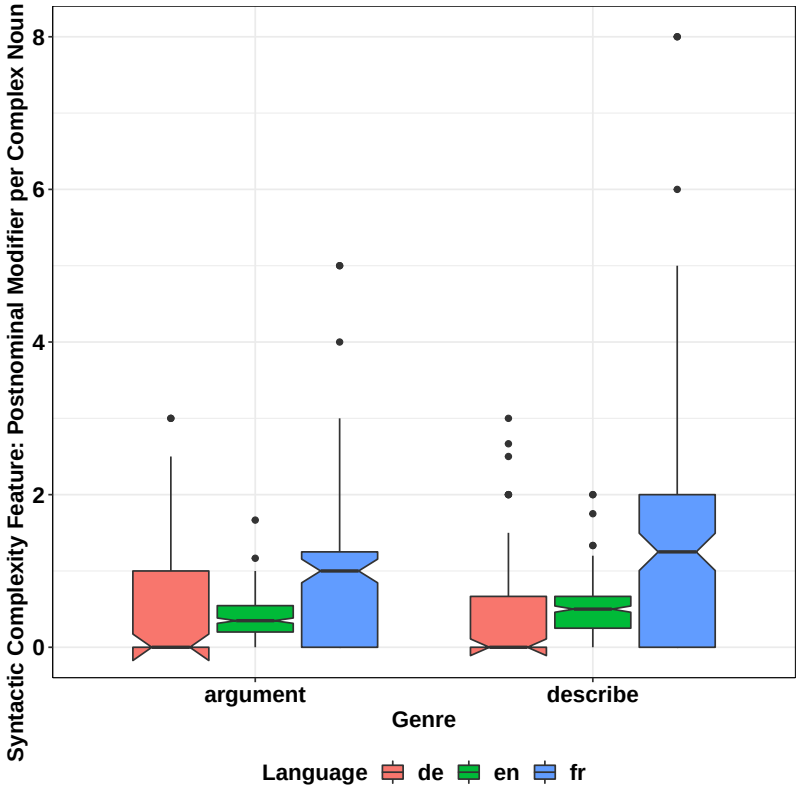
Noun phrase complexity



- Argumentative texts often found to contain more complex NPs than narrative texts in previous research (Yoon & Polio 2017b)
- English: argumentative texts > complex NP than descriptive texts
  - Effect absent in French and German argumentative texts

Pre- and postnominal modifiers





- Descriptive texts contain more pre- and postnominal modification across languages (effect weaker for German)
- Languages differ with regard to the amount of modification

Discussion

- Model informed by frequency, syntactic & lexical complexity
  - Descriptive: more variable noun use, rare words
  - Argumentative: more variable modifiers, clausal elaboration

Discussion

- Model informed by frequency, syntactic & lexical complexity
  - Descriptive: more variable noun use, rare words
  - Argumentative: more variable modifiers, clausal elaboration
- Uninformative: human processing, cohesion, & verb MCI
  - Matches Merlin/Falko findings for human processing
  - Cohesion: proficiency level or other task factors?

Discussion

- Model informed by frequency, syntactic & lexical complexity
  - Descriptive: more variable noun use, rare words
  - Argumentative: more variable modifiers, clausal elaboration
- Uninformative: human processing, cohesion, & verb MCI
  - Matches Merlin/Falko findings for human processing
  - Cohesion: proficiency level or other task factors?
- Results partially in line with research on genre differences in English between narrative and argumentative texts

Discussion

- Model informed by frequency, syntactic & lexical complexity
  - Descriptive: more variable noun use, rare words
  - Argumentative: more variable modifiers, clausal elaboration
- Uninformative: human processing, cohesion, & verb MCI
  - Matches Merlin/Falko findings for human processing
  - Cohesion: proficiency level or other task factors?
- Results partially in line with research on genre differences in English between narrative and argumentative texts
- Most features behave similar across languages despite presumed proficiency and language differences

Summary & outlook

1. Broad linguistic complexity modeling helps distinguish text genres for English, German, & French texts

Summary & outlook

1. Broad linguistic complexity modeling helps distinguish text genres for English, German, & French texts
2. Genre generalizes across two instances of languages: L2 & language of schooling

Summary & outlook

1. Broad linguistic complexity modeling helps distinguish text genres for English, German, & French texts
2. Genre generalizes across two instances of languages: L2 & language of schooling
3. Linguistic genre differences in lexicon, syntax and frequency, not in human processing and cohesion
   - Not fully in line with previous findings

⇒ Complex interplay of task factors, proficiency & complexity needs to be investigated in more detail

References

Biber, D. & S. Conrad (2019). *Register, genre, and style*. Cambridge University Press.

Brezina, V. & G. Pallotti (2019). Morphological complexity in written L2 texts. *Second language research* 35(1), 99–119.

Brysbaert, M., M. Buchmeier, M. Conrad, A. M. Jacobs, J. Bölte & A. Böhl (2011a). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology* 58, 412–424.

Brysbaert, M., E. Keuleers & B. New (2011b). Assessing the usefulness of Google Books' word frequencies for psycholinguistic research on word processing. *Frontiers in Psychology* 2(27).

Bulté, B. & A. Housen (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing* 26(0), 42 – 65. URL http://www.sciencedirect.com/science/article/pii/S1060374314000666. Comparing perspectives on L2 writing: Multiple analyses of a common corpus.

Chen, X. & D. Meurers (2016). CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*. Osaka, Japan: COLING, pp. 113–119.

Chen, X. & D. Meurers (2017). Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading* 41(3), 486–510.

Crossley, S. A., S. Skalicky & M. Dascalu (2019). Moving beyond classic readability formulas: new methods and new models. *Journal of Research in Reading* 42(3-4), 541–561.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita & W. O'Neil (eds.), *Image, language, brain: papers from the First Mind Articulation Project Symposium*, MIT, pp. 95–126.

Housen, A., F. Kuiken & I. Vedder (2012). Complexity, accuracy and fluency: Definitions, measurement and research. In A. Housen, F. Kuiken & I. Vedder (eds.), *Dimensions of L2 Performance and Proficiency*, John Benjamins, Language Learning & Language Teaching, pp. 1–20. URL https://doi.org/10.1075/lllt.32.01hou.

New, B., C. Pallier, L. Ferrand & R. Matos (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE™//A lexical database for contemporary french: LEXIQUE™. *L'année psychologique* 101(3), 447–462.

Pallotti, G. (2009). CAF: Defining, Refining and Differentiating Constructs. *Applied Linguistics* 30(4), 590–601. URL http://applij.oxfordjournals.org/content/30/4/590.full.pdf.

Robinson, P. (2001). Task complexity, task difficulty, and task production: exploring interactions in a componential framework. *Applied Linguistics* 22(1), 27–57.

Siemen, P., A. Lüdeling & F. H. Müller (2006). FALKO - ein fehlerannotiertes Lernerkorpus des Deutschen. In *Proceedings of Konvens*. Konstanz. URL http://www2.informatik.hu-berlin.de/~siemen/daten/falko.pdf.

Skehan, P. (1996). A Framework for the Implementation of Task-based Instruction. *Applied Linguistics* 17(1), 38.

Tracy-Ventura, N. & F. Myles (2015). The importance of task variability in the design of learner corpora for SLA research. *International Journal of Learner Corpus Research* 1(1), 58–95.

Vajjala, S. & D. Meurers (2012). On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. pp. 163–173. http://aclweb.org/anthology/W12-2019.pdf.

Weiss, Z., X. Chen & D. Meurers (2021). Using Broad Linguistic Complexity Modeling for Cross-Lingual Readability Assessment pp. 38–54.

Weiss, Z., K. Lange-Schubert, B. Geist & D. Meurers (2022). Sprachliche Komplexitä"t im Unterricht: Eine computerlinguistische Analyse der gesprochenen Sprache von Lehrenden und Lernenden im naturwissenschaftlichen Unterricht in der Primar- und Sekundarstufe. *Zeitschrift fü"r Germanistische Linguistik* .

Weiss, Z. & D. Meurers (2018). Modeling the Readability of German Targeting Adults and Children: An Empirically Broad Analysis and its Cross-Corpus Validation. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*. Santa Fe, New Mexico, USA. https://www.aclweb.org/anthology/C18-1026.

Weiss, Z. & D. Meurers (2019a). Analyzing Linguistic Complexity and Accuracy in Academic Language Development of German across Elementary and Secondary School. In *Proceedings of the 14th Workshop*

on Innovative Use of NLP for Building Educational Applications (BEA). Florence, Italy: Association for Computational Linguistics.

Weiss, Z. & D. Meurers (2019b). Broad Linguistic Modeling is Beneficial for German L2 Proficiency Assessment. In A. Abel, A. Glaznieks, V. Lyding & L. Nicolas (eds.), *Widening the Scope of Learner Corpus Research. Selected Papers from the Fourth Learner Corpus Research Conference*. Louvain-La-Neuve: Presses Universitaires de Louvain.

Weiss, Z. & D. Meurers (2021). Analyzing the linguistic complexity of German learner language in a reading comprehension task: Using proficiency classification to investigate short answer data, cross-data generalizability, and the impact of linguistic analysis quality. *International Journal of Learner Corpus Research* 7(1), 84–131.

Weiss, Z., A. Riemenschneider, P. Schröter & D. Meurers (2019). Computationally Modeling the Impact of Task-Appropriate Language Complexity and Accuracy on Human Grading of German Essays. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*. Florence, Italy.

Wisniewski, K., K. Schöne, L. Nicolas, C. Vettori, A. Boyd, D. Meurers, A. Abel & J. Hana (2013). MERLIN: An online trilingual learner corpus empirically grounding the european reference levels in authentic learner data. In *ICT for Language Learning 2013, Conference Proceedings, Florence, Italy. Libreriauniversitaria. it Edizioni*.

Yang, W., X. Lu & S. C. Weigle (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing* 28, 53–67. URL https://doi.org/10.1016/j.jslw.2015.02.002.

Yoon, H.-J. (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. *System* 66, 130–141.

Yoon, H.-J. & C. Polio (2017a). The Linguistic Development of Students of English as a Second Language in Two Written Genres. *TESOL Quarterly* pp. 275–301. URL https://doi.org/10.1002/tesq.296.

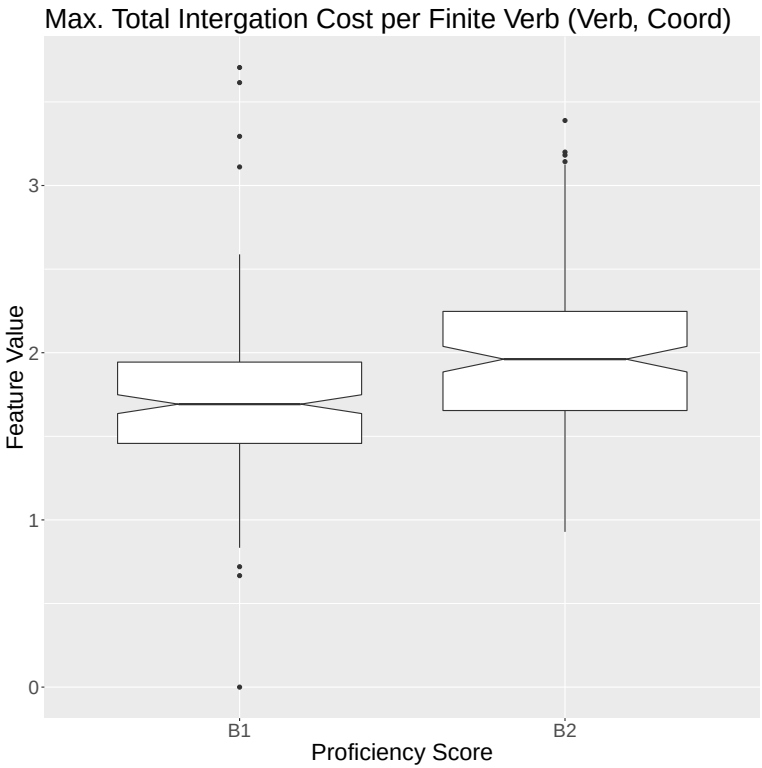
Yoon, H.-J. & C. Polio (2017b). The reliability and validity of automated tools for examining variation in syntactic complexity across genres. *Int J Appl Linguist* 28, 165–188.

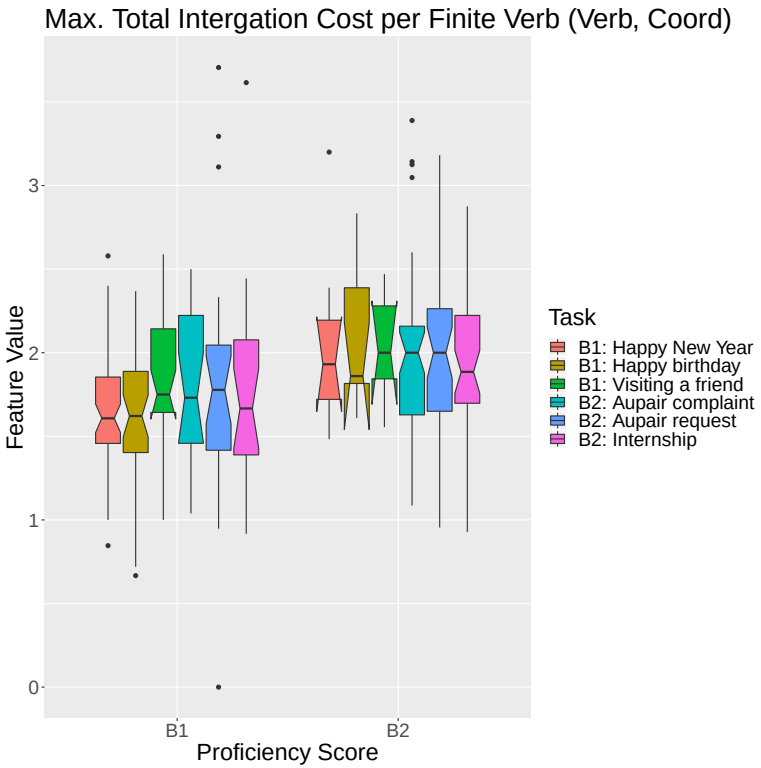
B1/B2 tasks in the Merlin corpus

		01	02	03	04	05	06
Test level	B1	✗	✗	✗			
	B2				✗	✗	✗
Genre	arg.				✗	✗	
	des.	✗	✗	✗			✗
Formality	form.				✗	✗	✗
	inform.	✗	✗	✗			
Structured/	yes	✗	✗	✗			
code complexity	no				✗	✗	✗

- B1: announce a visit, happy birthday, happy new year
- B2: Apply for job, work complaint, request job information

Human Processing Costs

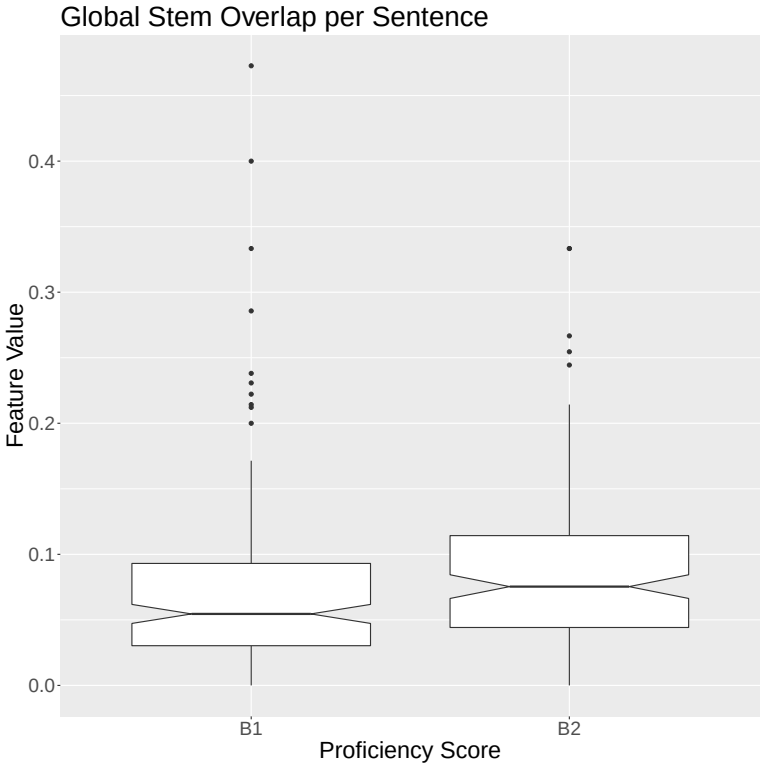


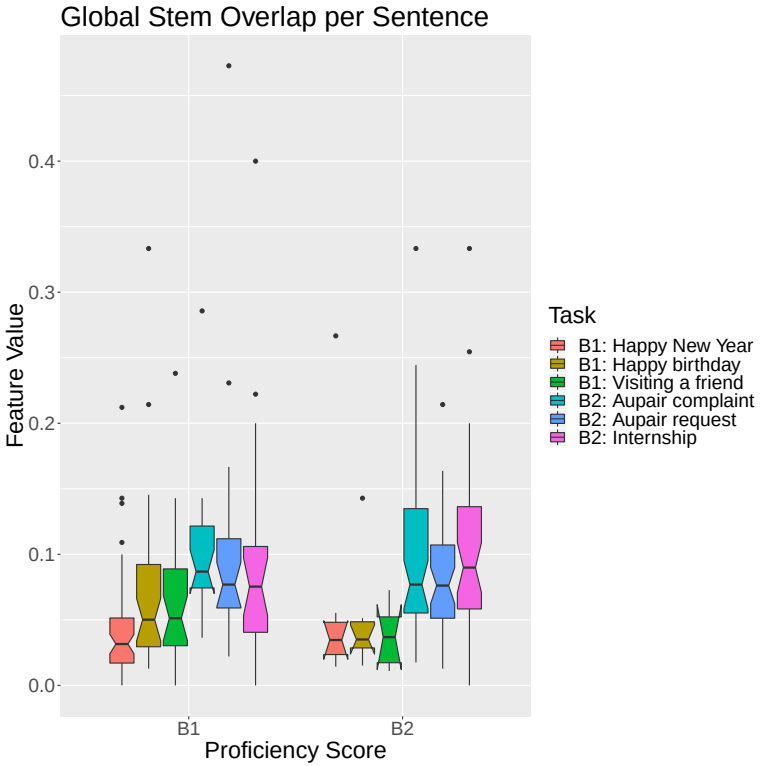


⇒ Integration costs are higher for texts with B2 scores

⇒ Withing proficiency levels, they are stable across tasks

Discourse Complexity





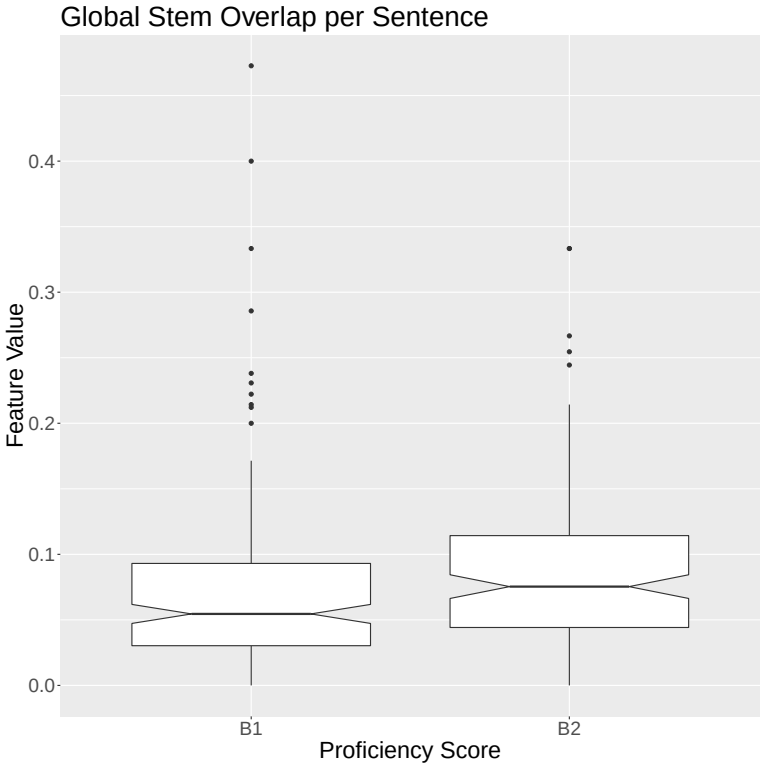
⇒ B2 test takers write more implicitly cohesive texts

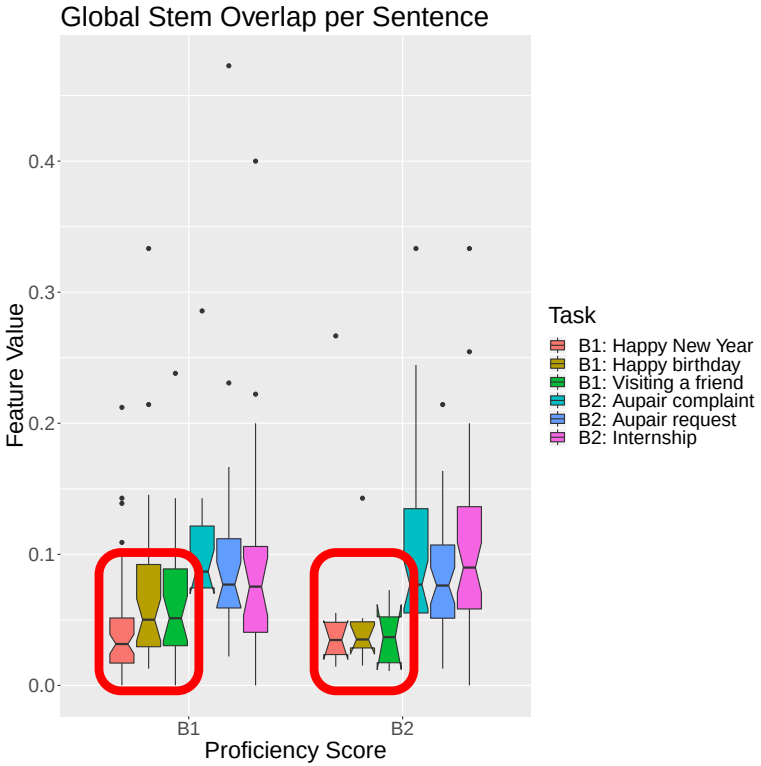
- Limited within task difference between B1 and B2 texts

⇒ Overall differences between B1 and B2 texts can be explained by the elicitation bias of CEFR testing

- B2 tests were more often taken by B2 than B1 learners

Discourse Complexity





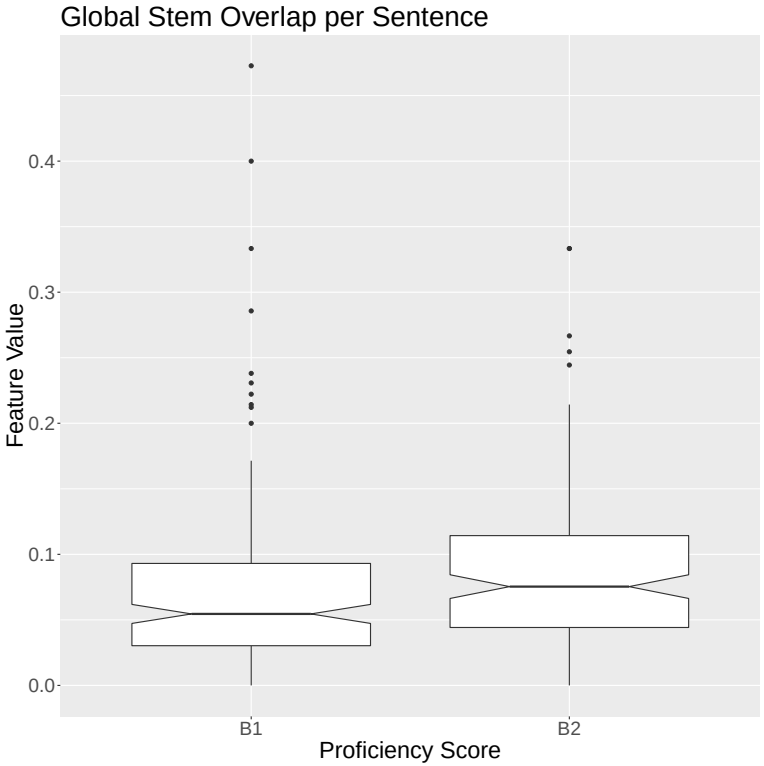
⇒ B2 test takers write more implicitly cohesive texts

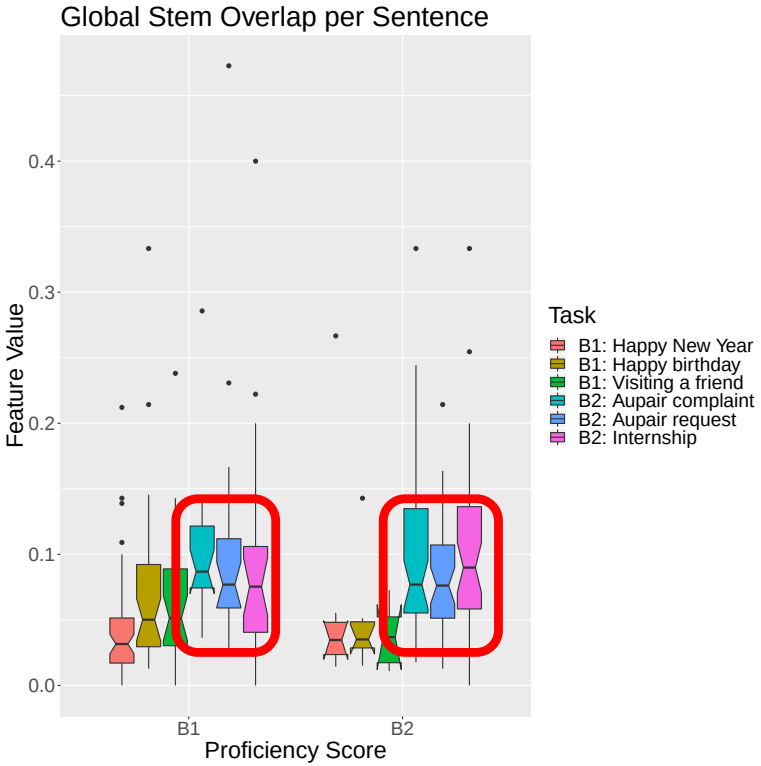
- Limited within task difference between B1 and B2 texts

⇒ Overall differences between B1 and B2 texts can be explained by the elicitation bias of CEFR testing

- B2 tests were more often taken by B2 than B1 learners

Discourse Complexity





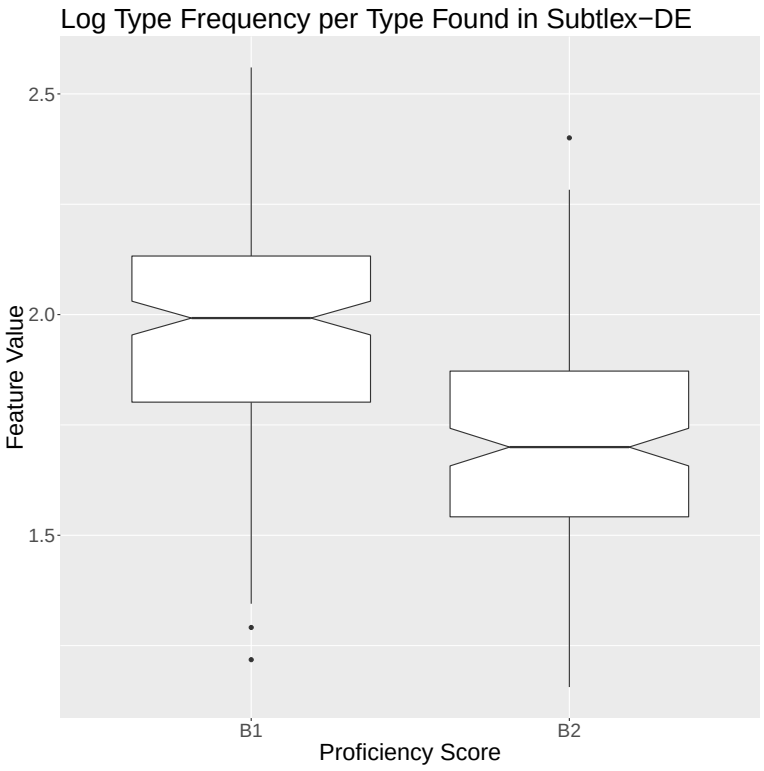
⇒ B2 test takers write more implicitly cohesive texts

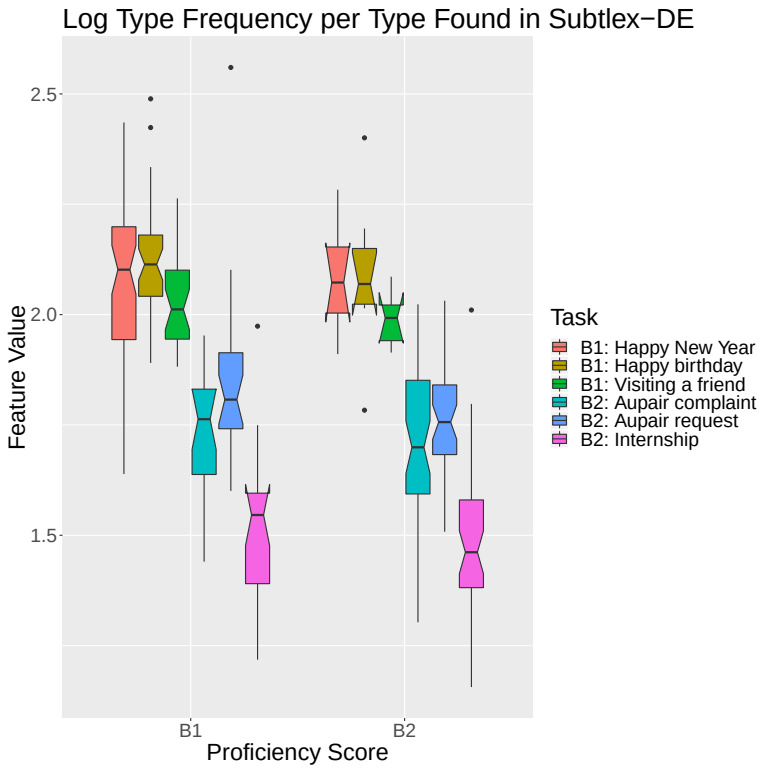
- Limited within task difference between B1 and B2 texts

⇒ Overall differences between B1 and B2 texts can be explained by the elicitation bias of CEFR testing

- B2 tests were more often taken by B2 than B1 learners

Language Use



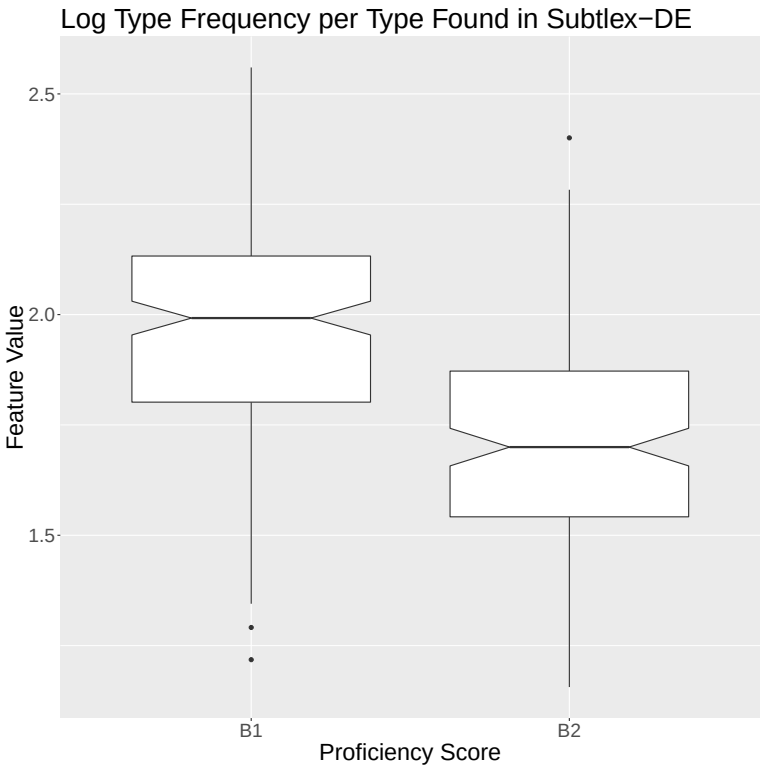


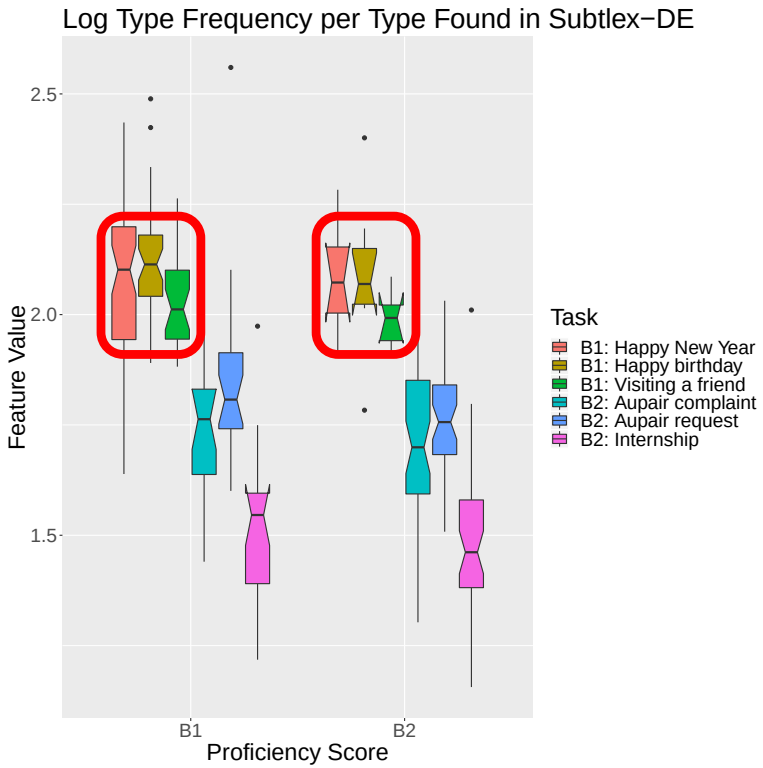
⇒ Tasks for B2 testing elicit more rare vocabulary

- Within tasks, texts by B1 and B2 learners do not differ
- This pattern holds for all frequency features we measured

⇒ Overall differences between B1 and B2 texts can be explained by the elicitation bias of CEFR testing

Language Use



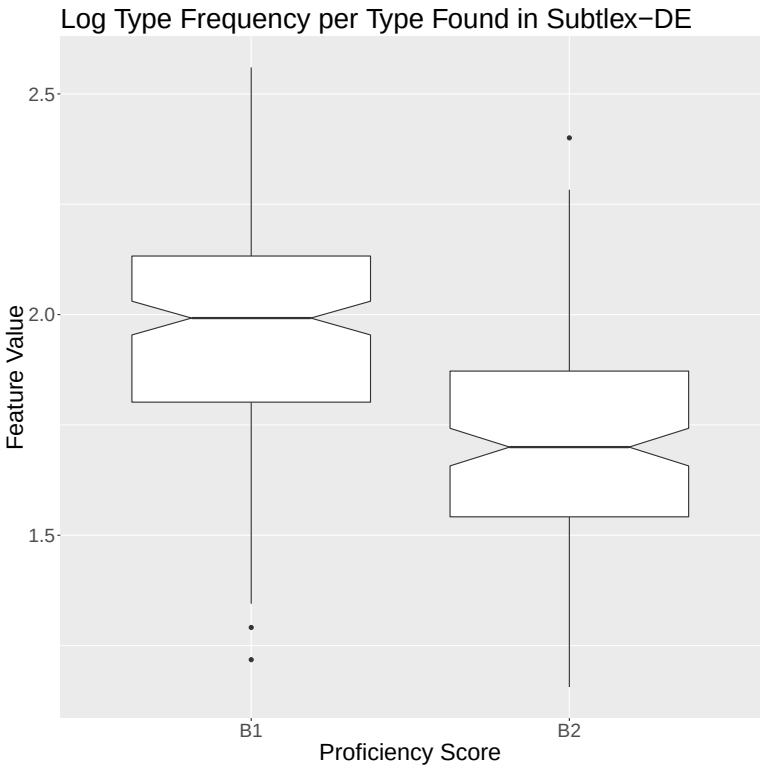


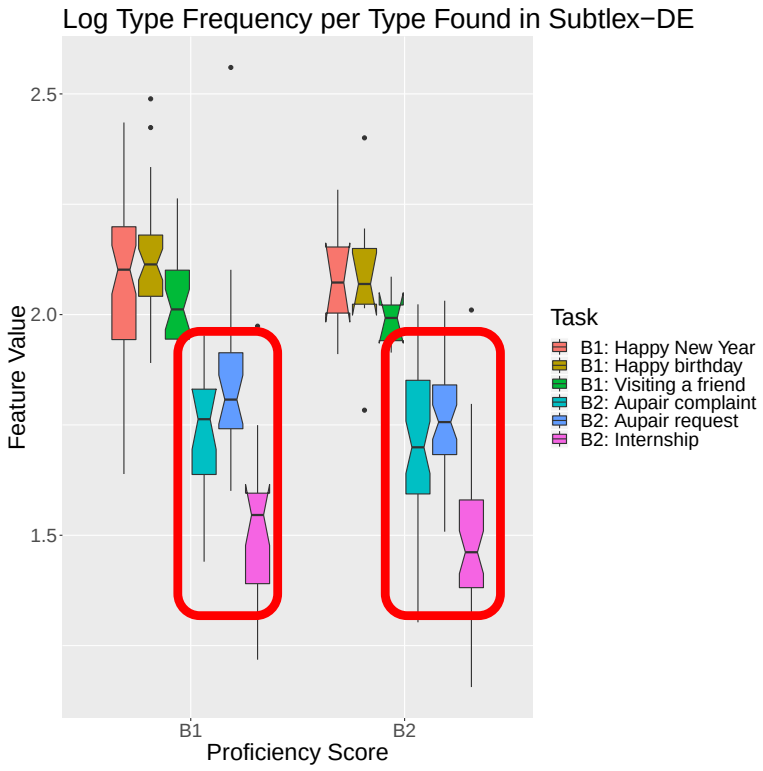
⇒ Tasks for B2 testing elicit more rare vocabulary

- Within tasks, texts by B1 and B2 learners do not differ
- This pattern holds for all frequency features we measured

⇒ Overall differences between B1 and B2 texts can be explained by the elicitation bias of CEFR testing

Language Use



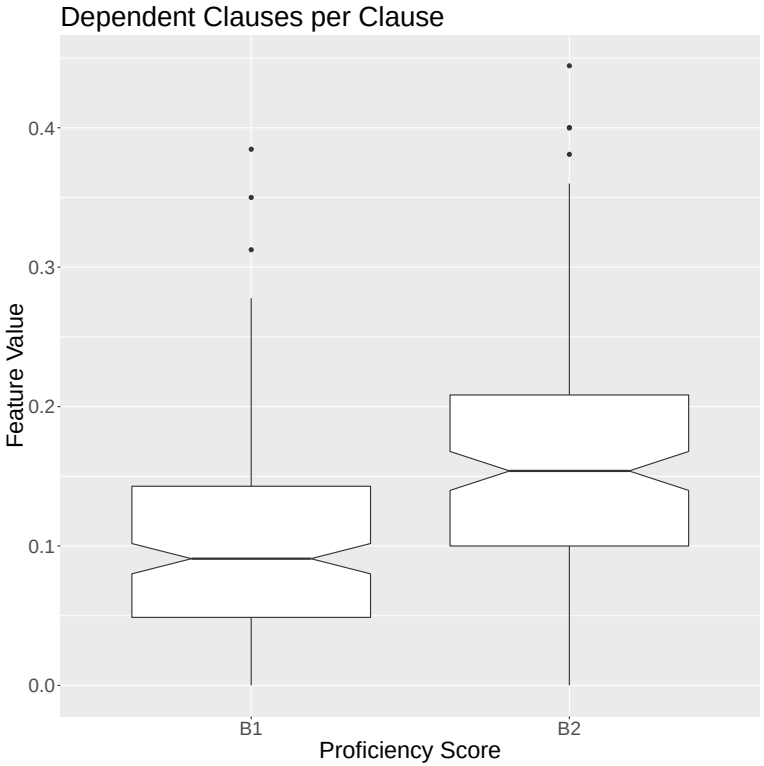


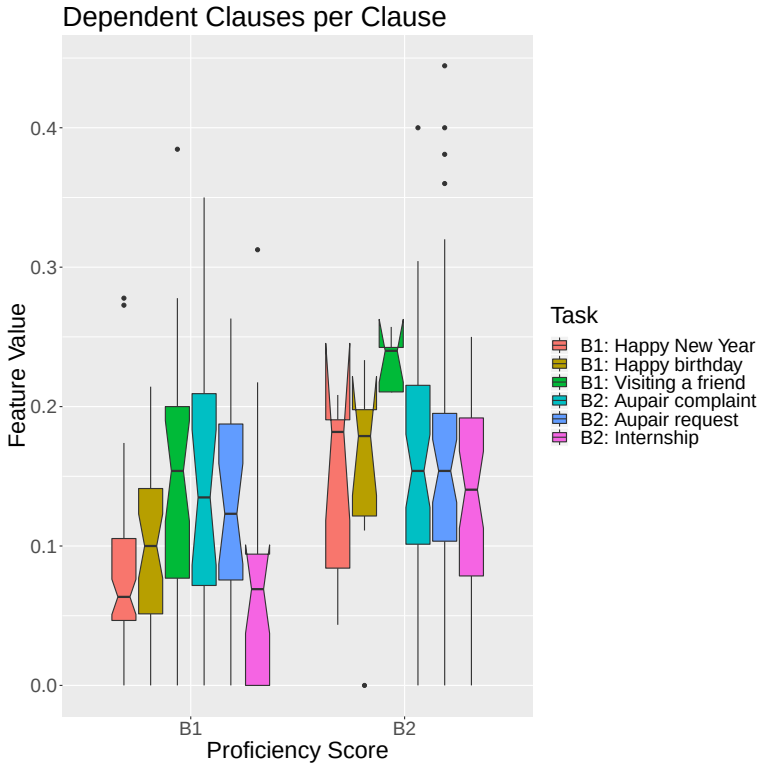
⇒ Tasks for B2 testing elicit more rare vocabulary

- Within tasks, texts by B1 and B2 learners do not differ
- This pattern holds for all frequency features we measured

⇒ Overall differences between B1 and B2 texts can be explained by the elicitation bias of CEFR testing

Clausal Complexity





⇒ B2 texts use more elaborate subordination than B1 texts

- This holds across measures of clausal complexity

⇒ We see some some task-specific variation at B1 and B2

- Specific tasks elicit more or less clausal elaboration
- This holds across measures of clausal complexity